KiDS-1000: Cosmic shear with enhanced redshift calibration

J. L. van den Busch1, A. H. Wright1, H. Hildebrandt1, M. Bilicki2, M. Asgari3,5, S. Joudaki6,7, C. Blake8, C. Heymans4,1, A. Kannawadi9, H. Y. Shan10,11, and T. Tröster4

1 Ruhr-University Bochum, Astronomical Institute, German Centre for Cosmological Lensing, Universitätsstr. 150, 44801 Bochum, Germany
e-mail: jlvdb@astro.rub.de
2 Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
3 Center for Theoretical Physics, Polish Academy of Sciences, al. Lotników 32/46, 02-668 Warsaw, Poland
4 Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
5 E. A. Milne Centre, University of Hull, Cottingham Road, Hull HU6 7RX, UK
6 Waterloo Centre for Astrophysics, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada
7 Department of Physics and Astronomy, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada
8 Centre for Astrophysics & Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia
9 Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA
10 Shanghai Astronomical Observatory (SHAO), Nandan Road 80, Shanghai 200030, PR China
11 University of Chinese Academy of Sciences, Beijing 100049, PR China

Received 24 August 2021 / Accepted 4 April 2022

ABSTRACT

We present a cosmic shear analysis with an improved redshift calibration for the fourth data release of the Kilo-Degree Survey (KiDS-1000) using self-organising maps (SOMs). Compared to the previous analysis of the KiDS-1000 data, we expand the redshift calibration sample to more than twice its size, now consisting of data of 17 spectroscopic redshift campaigns, and significantly extending the fraction of KiDS galaxies we are able to calibrate with our SOM redshift methodology. We then enhanced the calibration sample with precision photometric redshifts from COSMOS2015 and the Physics of the Accelerated Universe Survey (PAUS), allowing us to fill gaps in the spectroscopic coverage of the KiDS data. Finally we performed a Complete Orthogonal Sets of E/B-Integrals (COSEBIs) cosmic shear analysis of the newly calibrated KiDS sample. We found $\chi^2_8 = 0.748^{+0.021}_{-0.020}$, which is in good agreement with previous KiDS studies and increases the tension with measurements of the cosmic microwave background to 3.4$\sigma$. We repeated the redshift calibration with different subsets of the full calibration sample and obtained, in all cases, agreement within at most 0.5$\sigma$ in $\chi^2$ compared to our fiducial analysis. Including additional photometric redshifts allowed us to calibrate an additional 6% of the source galaxy sample. Even though further systematic testing with simulated data is necessary to quantify the impact of redshift outliers, precision photometric redshifts can be beneficial at high redshifts and to mitigate selection effects commonly found in spectroscopically selected calibration samples.

Key words. cosmology: observations – gravitational lensing: weak – galaxies: distances and redshifts – surveys

1. Introduction

Over the past decade, gravitational lensing (Bartelmann & Schneider 2001) has emerged as one of the most powerful tools to study gravity and the dark sectors of the Universe, dark matter and dark energy, through the impact of these components on the density fluctuations of matter and their evolution with cosmic time (Peacock et al. 2006). In the limit of weak lensing, massive structures along the line of sight imprint a subtle shearing on the shapes of distant galaxies. This signal can be extracted by statistically analysing the ellipticity of galaxy images in large surveys (Refregier 2003). These cosmic shear surveys (Kilbinger 2015) face the challenge that they must accurately reconstruct the galaxy redshift distribution in order to interpret the cosmological signal correctly. Even small biases in the first moment of the redshift distribution may introduce significant biases in the recovered cosmological parameters (e.g. Huterer et al. 2006; Ma et al. 2006). There exists tension between constraints from cosmic shear and the cosmic microwave background, first seen between the Canada-France-Hawaii Telescope Lensing Survey (CFHTLenS, Heymans et al. 2013; MacCrann et al. 2015; Joudaki et al. 2017) and Planck (Planck Collaboration XVI 2014), but also between the Kilo-Degree Survey (KiDS, Kuijken et al. 2015) and Planck legacy (Planck Collaboration VI 2020) for example, and recently for KiDS-1000 (Asgari et al. 2021). The most recent cosmic shear results from the Dark Energy Survey (DES, Flaugher et al. 2015) are very similar to KiDS-1000 (Amon et al. 2022; Secco et al. 2022), albeit at a lower statistical tension with Planck1. In the light of these repeatedly reported tensions, redshift calibration has come under scrutiny as one of the systematics for cosmic shear experiments (e.g. Joudaki et al. 2020).

Due to the statistical nature of the shear measurements, current generation (stage-III) cosmic shear surveys, such as KiDS, DES, and the Hyper Suprime-Cam Subaru Strategic Program (HSC, Aihara et al. 2018), rely on the imaging of tens of millions of galaxies for which spectroscopic redshifts cannot be measured

1 Conclusions on the degree of tension differ primarily owing to the way tension is quantified and different prior choices for the neutrino mass in the Planck reanalysis.

Open Access article, published by EDP Sciences, under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article is published in open access under the Subscribe-to-Open model. Subscribe to A&A to support open access publication.
directly. Instead, galaxy redshifts are determined with secondary redshift estimates, the most notable ones are direct calibration with spectroscopic training samples (e.g. Lima et al. 2008; Hildebrandt et al. 2017, 2020; Buchs et al. 2019; Wright et al. 2020a), clustering redshifts (which infer redshift distributions by exploiting the gravitational clustering of galaxies at similar redshifts, e.g. Newman 2008; Matthews & Newman 2010; Schmidt et al. 2013; Ménard et al. 2013; van den Busch et al. 2020; Hildebrandt et al. 2021; Gatti et al. 2022), and methods that make use of a combination of both these approaches (Sánchez & Bernstein 2019; Alarcon et al. 2020; Myles et al. 2021).

The redshift calibration of the fourth data-release of KiDS (Kuijken et al. 2019; Hildebrandt et al. 2021) relies on an implementation of the direct calibration that utilises a self-organising map (SOM, Kohonen 1982; Wright et al. 2020a) based on work by Masters et al. (2016). The fundamental principle of this method is to re-weight a spectroscopic reference sample such that it is representative of a photometric dataset with six new redshift distribution. The weighted redshift distribution of the reference sample is then a direct estimate of the unknown distribution. Additionally, the SOM method allows for the removal of galaxies from the photometric dataset for which no similar reference sample is then a direct estimate of the unknown distribution. This work is based on the fourth data release of KiDS which covers 1006 deg². The weak lensing source catalogue (KiDS-1000, Giblin et al. 2021) is divided into five tomographic redshift bins, based on nine-band photometric redshifts (four bins with ΔzB = 0.2, starting from zB = 0.1, and a fifth bin at 0.9 < zB ≤ 1.2) computed by BPZ (Bayesian Photometric Redshift, Benítez 2000). It contains all objects with non-zero shear measurement weights obtained from lensfit (Miller et al. 2007; Fenech Conti et al. 2017), which effectively selects objects with r-band magnitudes between 20 ≤ r ≤ 25.

In addition to the main survey imaging there are observations of six fields dedicated primarily to redshift calibration. These ‘KiDZ’ fields cover approximately 1 deg² each and target areas of the sky also observed by different spectroscopic campaigns, which are summarised in Sect. 2.2. Just like the main survey, KiDZ provides KiDS+ VIKING nine-band imaging which reaches or exceeds the depth of the main survey. In the latter case we homogenise the data depth by applying Gaussian noise to obtain matched photometry.

2.1. KiDS-1000 photometric data

The Kilo-Degree Survey (KiDS, Kuijken et al. 2015; de Jong et al. 2015, 2017; Kuijken et al. 2019) is a public European Southern Observatory (ESO) survey that has been designed particularly with weak gravitational lensing applications in mind. The complete survey will deliver about 1350 deg² of ugri imaging split into an equatorial and a southern field. Combined with ZYJHK, imaging from its companion infrared survey, the VISTA Kilo-Degree Infrared Galaxy Survey (VIKING, Edge et al. 2013; Venemans et al. 2015), this constitutes a nine-band, matched-depth data-set with primary imaging in the r-band, observed at a mean seeing of 0.7''. This work is based on the fourth data release of KiDS which covers 1006 deg². The weak lensing source catalogue (KiDS-1000, Giblin et al. 2021) is divided into five tomographic redshift bins, based on nine-band photometric redshifts (four bins with ΔzB = 0.2, starting from zB = 0.1, and a fifth bin at 0.9 < zB ≤ 1.2) computed by BPZ (Bayesian Photometric Redshift, Benítez 2000). It contains all objects with non-zero shear measurement weights obtained from lensfit (Miller et al. 2007; Fenech Conti et al. 2017), which effectively selects objects with r-band magnitudes between 20 ≤ r ≤ 25.

In addition to the main survey imaging there are observations of six fields dedicated primarily to redshift calibration. These ‘KiDZ’ fields cover approximately 1 deg² each and target areas of the sky also observed by different spectroscopic campaigns, which are summarised in Sect. 2.2. Just like the main survey, KiDZ provides KiDS+ VIKING nine-band imaging which reaches or exceeds the depth of the main survey. In the latter case we homogenise the data depth by applying Gaussian noise to obtain matched photometry.

2.2. Spectroscopic data for calibration

The most important spectroscopic campaigns that overlap with the six KiDZ fields are zCOSMOS (Lilly et al. 2009), VVDS (VIMOS VLT Deep Survey, Le Fèvre et al. 2005, 2013), DEEP2 (Newman et al. 2013), the GAMA (Galaxy And Mass Assembly, Driver et al. 2011) deep field G15deep (Kafle et al. 2018; Driver et al. 2022), and a compilation of spectra covering the Chandra Deep Field South (CDF-S, Appendix A). This data has been used in previous KiDS redshift calibration works (Wright et al. 2020a; Hildebrandt et al. 2021).

We extended this compilation by adding data from C3R2 (Complete Calibration of the Colour-Redshift Relation, Masters et al. 2017, 2019; Euclid Collaboration 2020; Stanford et al. 2021), DEVILS (Deep Extragalactic Visible Legacy Survey, Davies et al. 2018), VIPERS (VIMOS 3 Public Extragalactic Redshift Survey, Scodégio et al. 2018), and a variety of spectroscopic campaigns that target the CDF-S and COSMOS fields which are detailed in Appendix A. We also revised the selection of sources included for calibration by removing duplicates, both from spatial overlap as well as within the datasets, and by homogenising redshift quality flags based on the original information in the input samples. If, for a given source, there are redshifts from different surveys available, we assigned the most reliable measurement based on a specific

in 2021

In case of the COSMOS field we instead use existing data in the CDF-S, Appendix A. This data has been used in previous KiDS redshift calibration works (Wright et al. 2020a; Hildebrandt et al. 2021).

We extended this compilation by adding data from C3R2 (Complete Calibration of the Colour-Redshift Relation, Masters et al. 2017, 2019; Euclid Collaboration 2020; Stanford et al. 2021). We added the following data sets to the compilation in Appendix A: C3R2 (Complete Calibration of the Colour-Redshift Relation, Masters et al. 2017, 2019; Euclid Collaboration 2020; Stanford et al. 2021), DEVILS (Deep Extragalactic Visible Legacy Survey, Davies et al. 2018), VIPERS (VIMOS 3 Public Extragalactic Redshift Survey, Scodégio et al. 2018), and a variety of spectroscopic campaigns that target the CDF-S and COSMOS fields which are detailed in Appendix A. We also revised the selection of sources included for calibration by removing duplicates, both from spatial overlap as well as within the datasets, and by homogenising redshift quality flags based on the original information in the input samples. If, for a given source, there are redshifts from different surveys available, we assigned the most reliable measurement based on a specific
specific ‘hierarchy’ of surveys (see Appendix A for details). For objects with multiple spectroscopic measurements within a particular survey, we either took the redshift with the highest quality flag or, if various entries for the same source have the same quality flag and the reported redshifts differ by no more than 0.005, we took the average. However, if the reported redshift differences exceed this threshold, we excluded such a source from the compilation. We restricted the selection to objects with high quality spectroscopic redshifts (approximately corresponding to ≥95% confidence or redshift quality code nQ ≥ 3). Figure 1 compares the number of galaxies and their mean redshift for all samples that enter the spectroscopic compilation. These values apply after removing duplicates between overlapping catalogues and only for those objects with photometric coverage in KiDS (Sect. 2.1).

2.3. Photometric data for calibration

The success rate of spectroscopically determined redshifts is very different from (typically flux-limited) imaging data. Therefore, it is very difficult to obtain a spectroscopic calibration sample that is representative of photometric data in magnitude and colour space, especially at faint magnitudes. Instead, we additionally included galaxy samples with high quality photometric redshifts to achieve a greater overall coverage of the KiDS data by the calibration sample which was beneficial for our redshift calibration technique of choice (Sect. 3.1).

2.3.1. COSMOS2015

The COSMOS2015 catalogue (Laigle et al. 2016) constitutes a sample of about half a million galaxies in the COSMOS field with precision photometric redshifts derived from up to 30 photometric bands, ranging from near ultra-violet to mid infrared, including 14 medium and narrow band filters. This sample extends to higher redshifts \( z_{\text{max}} \approx 6 \) and fainter magnitudes than our spectroscopic compilation, but at the cost of less secure redshift estimates with an outlier fraction ranging from 0.5% at low redshifts to 13.2% for \( 3 < z < 6 \) (Laigle et al. 2016).

2.3.2. PAUS

The PAUS photometric redshift sample \(^5\) (Alarcon et al. 2021) is a combination of 26 optical and near-infrared bands from COSMOS2015 that are matched against observations of the COSMOS field in 40 narrow band filters by the PAU survey. These PAU filters sample the optical regime between 450 nm to 850 nm at a bandwidth of \( \Delta \lambda = 12.5 \text{ nm} \) (Padilla et al. 2019) and the combined photometric catalogue is limited to \( i_{\text{AB}} < 23 \). Due to the relatively high spectral resolution of the dataset a new Bayesian spectral energy distribution (SED) fitting technique is required that accounts for individual emission lines (Alarcon et al. 2021). This allows the PAUS photo-\( \cdot \)z to achieve a \( 3\times(1.7\times) \) lower photo-\( \cdot \)z scatter at the bright (faint) end of the magnitude distribution and marginally smaller outlier fractions compared to the original COSMOS2015 photo-\( \cdot \)z at \( i_{\text{AB}} < 23 \). The scaled photo-\( \cdot \)z bias is very low and has a \( \{\text{median}(\Delta z)\} < 0.001 \) over the whole redshift range of the PAUS sample. Therefore this sample positions itself right between the spectroscopic data and COSMOS2015 in terms of completeness and redshift precision.

2.4. Combined calibration sample

In this work we selected data from the full COSMOS2015 photometric catalogue and combined this data hierarchically with PAUS and the spectroscopic data. Finally we matched this unified catalogue to the KiDS imaging to form our redshift calibration sample.

We prepared the full COSMOS2015 photometric catalogue similar to Laigle et al. (2016), that is, we selected only those sources which fall into the intersection of the footprint of the COSMOS field (\( \text{flag}_{\text{cosmos}} = 1 \)) and the UltraVISTA observations (\( \text{flag}_{\text{hjmcx}} = 0 \)), which provide essential infrared spectral coverage. We excluded data from saturated areas (\( \text{flag}_{\text{peter}} = 0 \)) and additionally removed objects that are classified to be most likely stars (\( \text{type} \neq 1 \)) or have no photo-\( \cdot \)z estimate. This selection yields about half a million objects.

About 40 000 sources of the PAUS sample are, by design, matched against COSMOS2015 and therefore require no further preparation. Therefore we were able to directly combine the spectroscopic compilation, PAUS, and the subset of COSMOS2015 by matching objects within \( 1'' \). We maintained a hierarchy to ensure that we always chose the most reliable redshift estimate available: spec-\( \cdot \)z supersedes PAUS photo-\( \cdot \)z which supersedes the COSMOS2015 photo-\( \cdot \)z. Finally, we assigned nine-band KiDS magnitudes to this compilation by matching against the KiDS data, again within \( 1'' \). This combination of spec-\( \cdot \)z (in all KiDS fields), photo-\( \cdot \)z (only in COSMOS), and KiDS imaging represents our full redshift calibration sample.

The method to combine the two photometric redshift samples with our spectroscopic compilation in an hierarchical manner is very similar to the approach taken for the redshift calibration of the DES Y3 data (Myles et al. 2021). There are, however, two key differences to their approach. First, our compilation of spectroscopic redshifts covers a much wider range of the colour-redshift space than their selection of spectra. This allowed us to construct more representative calibration samples that consist purely of spectroscopic redshifts, photometric redshifts, or a

\(^4\) We used the median of the photo-\( \cdot \)z likelihood distribution (PHOTOZ column in the catalogue).

\(^5\) Available at cosmohub.pic.es
Table 1. Number counts and mean redshifts of the original KiDS-1000 redshift calibration sample and different subsets of the new redshift compilation.

<table>
<thead>
<tr>
<th>Compilation</th>
<th>Count</th>
<th>\langle z \rangle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hildebrandt et al. (2021)</td>
<td>25,373</td>
<td>0.796</td>
</tr>
<tr>
<td>spec-z fiducial</td>
<td>52,911</td>
<td>0.788</td>
</tr>
<tr>
<td>spec-z+PAUS</td>
<td>61,163</td>
<td>0.776</td>
</tr>
<tr>
<td>spec-z+PAUS+COS15</td>
<td>112,400</td>
<td>1.002</td>
</tr>
<tr>
<td>spec-z nQ ≥ 4</td>
<td>24,117</td>
<td>0.832</td>
</tr>
<tr>
<td>Only PAUS</td>
<td>24,229</td>
<td>0.640</td>
</tr>
<tr>
<td>Only COS15</td>
<td>80,632</td>
<td>1.081</td>
</tr>
<tr>
<td>Only PAUS+COS15</td>
<td>80,635</td>
<td>1.084</td>
</tr>
</tbody>
</table>

Notes. The full compilation is represented as spec-z+PAUS+COS15. Values apply after matching with the KiDS imaging and removing duplicates in the spectroscopic data.

combination thereof. Second, the primary KiDS imaging data and the calibration data are observed in all nine photometric bands at a comparable depth, which simplified the mapping from galaxy colour to redshift significantly (see Sect. 3).

2.4.1. Primary compilations

From this heterogeneous sample with redshift estimates from very different sources we selected three subsets, each with a higher redshift precision but lower completeness: Firstly the full compilation (to which we refer as spec-z+PAUS+COS15), secondly objects with either spectroscopic redshifts or PAUS photo-z (spec-z+PAUS), and finally our fiducial sample containing only those objects that have spectroscopic redshifts (spec-z fiducial). The main properties and redshift distributions of these three primary compilations are summarised in Table 1 and Fig. 2.

The fiducial sample is already about twice as large as the calibration sample used previously by Hildebrandt et al. (2021) to calibrate the KiDS-1000 redshifts. Of the additional spectra we consider DEVILS and C3R2, the latter of which is designed to target regions of the galaxy colour-space with currently little spectroscopic coverage, to be the most important contributors. Similarly spec-z fiducial already contains about 66% of the matched PAUS sources, of which the majority has redshift $z < 1$. Due to its limited depth, the spec-z+PAUS sample presents only a small improvement over the fiducial case. The COSMOS2015 data, on the contrary, nearly doubles the compilation to its final size of about 112,000 objects. Due to the significantly higher depth of the COSMOS2015 photo-z, the fraction of sources with $z > 1$ nearly triples, pushing the mean redshift to \( \langle z \rangle \approx 1.0 \). Nevertheless this comes at the cost of a lower redshift accuracy compared to the rest of the sample.

2.4.2. Secondary compilations

In addition to the three primary compilations we also considered a subset that is restricted to only the most secure spectroscopic redshifts. This spec-z nQ ≥ 4 sample is, due to the large fraction of shared spectra, closest to the one of Hildebrandt et al. (2021) except that it lacks some low redshift sources (Fig. 2).

Finally, we created three subsets of the full redshift compilation that rely purely on photometric redshift estimates. We achieved this by recompiling the redshift compilation according to Sect. 2.4 but omitted all spectroscopic redshifts, therefore maintaining the usual hierarchy of PAUS and COSMOS2015 photo-z. These are objects from only the PAU survey (only-PAUS), all objects with photo-z (only-PAUS+COS15), and the pure COSMOS2015 subset (only-COS15, also discarding PAUS photo-z from the stack). Since the PAUS sample is essentially a subset of the COSMOS2015 catalogue, the latter two samples are almost identical except that for 30% of the sources the photo-z are augmented by the PAUS data (see Table 1). The PAUS sample is about half the size of the fiducial spectroscopic compilation and, while achieving a higher completeness at $i_{AB} < 23$, lacks many important faint, high redshift objects.

3. Redshift calibration with self-organising maps

A self-organising map (SOM, Kohonen 1982) is a very powerful tool that allows us to calibrate the redshift distribution of the KiDS-1000 lensing sample using the redshift compilations defined in the previous section. We adopted the SOM methodology of Wright et al. (2020a) which additionally provides a metric to select only those parts of the KiDS colour space in which we can reliably map out the colour-redshift relation.

3.1. SOM redshift calibration methodology

The basic idea of the SOM methodology dates back to Lima et al. (2008) who introduced a redshift calibration strategy built on the assumption that two galaxy samples with the same colour-space distribution follow the same redshift distribution. Therefore they suggested to derive the unknown redshift distribution $N(z)$ of a photometric galaxy sample from a calibration sample with accurate, preferentially spectroscopic redshifts $N^{\text{cal}}(z)$ that is constructed such that it is representative of the photometric sample. This method is called ‘direct calibration’ (DIR). In practice, however, such a calibration sample has typically a substantially different selection function. Therefore Lima et al. (2008) proposed a re-weighting scheme to match the calibration to the photometric sample by computing the ratio of the local galaxy density of both samples in the high-dimensional colour-space spanned by the photometric observations. This can be achieved for example by counting neighbours in a fixed volume around a point in the colour-space or by computing the volume occupied by a fixed number of nearest neighbours. Provided that both samples initially cover the same volume of the
colour-space this method should recover the true redshift distribution, even in the presence of colour-redshift degeneracies.

This method is still susceptible in particular to selection biases and incompleteness introduced by spectroscopic targeting strategies and success rates (e.g. Gruen & Brimioulle 2017; Hartley et al. 2020). Recent work by Wright et al. (2020a) shows that this can be alleviated by performing additional cleaning and selections (quality control, see Sect. 3.2) on the unknown sample, creating a gold sample containing only galaxies of the photometric sample that are sufficiently represented by the calibration sample. They implement this by training a SOM on the colour-space of the calibration sample and then parse the photometric sample into the same cells. Cells that are not occupied by objects from both samples are rejected, effectively removing those critical parts of the colour-space. They improve the cleaning procedure by applying hierarchical clustering on the SOM to find groups of cells with similar photometric properties instead of filtering individual cells. This allows for a more fine-grained trade-off between the number of photometric sources rejected due to partitioning of the high-dimensional colour-space and the bias introduced by misrepresentation of the gold sample.

Finally, they compute the DIR weight for each of the n SOM groupings $G = \{g_1, \ldots, g_n\}$ which is the ratio of calibration-to-gold sample objects. They obtain the redshift distribution of the gold sample by calculating the DIR-weighted sum of the redshift distributions $N^\text{cal}_g(z)$ of the calibration sample in each SOM grouping. $N^\text{tot}_g$ and $N^\text{gold}_g$ are the total number of calibration sample and gold sample objects of group $g$, respectively.

$$N(z) = \sum_{g \in G} \frac{N^\text{cal}_g(z) \times N^\text{gold}_g}{N^\text{tot}_g} \quad (1)$$

3.2. Application to KiDS-1000

For our analysis we largely followed Wright et al. (2020a) and trained a SOM with 101 × 100 hexagonal cells and periodic boundaries on the full calibration sample (spec-z+PAUS+ COS15, see Sect. 2.4.1). The input features were the matched KiDS r-band magnitudes and all 36 possible KiDS-colours that can be formed from the $ugriZYJHK_s$ imaging. Next, we divided the calibration and the KiDS-1000 source sample into the five tomographic bins and parsed both samples into the SOM cells. We then ran the hierarchical clustering for which we used the same number of clusters per bin (4000, 2200, 2800, 4200, and 2000) as Wright et al. (2020a) since these numbers were calibrated using simulations6 (van den Busch et al. 2020). Even though each gold sample has a different optimal number of clusters, simulating the new redshift compilation and including realistic photo-z errors is beyond the scope of this work.

We used the same SOM for the remaining calibration samples defined in Sect. 2.4 and simply parsed the corresponding subset of the full calibration sample back into the SOM before running the hierarchical clustering. For each of these calibration samples we applied a final cleaning step to the SOM groupings by defining a quality cut

$$|\langle z_{\text{cal}} \rangle - \langle z_\text{b} \rangle| > 5\sigma_{\text{mad}}, \quad (2)$$

where $\sigma_{\text{mad}} = n\text{MAD}(\langle z_{\text{cal}} \rangle - \langle z_\text{b} \rangle)$ is the normalised median absolute deviation from the median, where the normalisation ensures that the nMAD reproduces the traditional standard deviation in the limit of Gaussian noise. This selection rejects clusters of SOM cells in which the mean calibration sample redshift $\langle z_{\text{cal}} \rangle$ and the mean KiDS photometric redshifts $\langle z_\text{b} \rangle$ catastrophically disagree. Wright et al. (2020a) find that this additional cleaning significantly reduces the SOM redshift bias while the impact on the number density is small and does not exceed a few percent. The rejection threshold of $\sigma_{\text{mad}} \approx 0.12$ was calculated for the spec-z fiducial case and was applied to all other samples. This choice was motivated by the fact that this value is very close to the one calibrated with mock data for KiDS-1000 by (Hildebrandt et al. 2021), whereas $\sigma_{\text{mad}}$ would have been twice as large if we had calculated this threshold based on spec-z+PAUS+COS15. One reason for this difference in $\sigma_{\text{mad}}$ is that the COSMOS2015 data allow the inclusion of additional populations of faint galaxies for which the calibration sample reference redshifts and the KiDS photo-z are more likely discrepant, increasing the spread of the distribution of $\langle z_{\text{cal}} \rangle - \langle z_\text{b} \rangle$. We discuss this effect further in Sect. 6.1 and Appendix B.

This final selection step defines our gold sample for which we computed the redshift distributions according to Eq. (1). Since we required weighted redshift distributions for our cosmological analysis we substituted $N^\text{gold}_{g,\text{tot}}$ by $w^\text{gold}_{g,\text{tot}} = \sum_{i \in G} w_i$, which is the sum over the individual galaxy weights $w_i$ from shape measurements in the SOM group $g$.

3.3. Clustering redshifts

There is one key difference to the calibration methodology of Hildebrandt et al. (2021) which is that we chose to omit the clustering redshift analysis in this work. While this choice limits our ability to validate the redshift distributions of our new gold samples, the SOM method is our fiducial calibration method and is therefore the focus of this work. In addition to that, the newly included calibration data does not increase the spatial overlap with KiDS significantly and is, due to its inhomogeneity, difficult to administer in a cross-correlation analysis. We leave this validation and joint analysis with the clustering redshifts to future work.

4. Cosmological analysis

In this section we summarise our cosmic shear analysis pipeline which we adopt from Asgari et al. (2021, A21 hereafter).

4.1. Cosmic shear

The primary observable of cosmic shear are the shear two-point correlation functions (2PCFs, Kaiser 1992)

$$\xi_{ij}(\theta) = \langle \gamma_i \gamma_j(\theta) \rangle - \langle \gamma_i \rangle \langle \gamma_j \rangle, \quad (3)$$

where $\gamma_i$ and $\gamma_j$ are the tangential and the cross component of the shear, defined with respect to the line connecting a pair of galaxies (see e.g. Bartelmann & Schneider 2001). We used a weighted estimator for the shear correlations $\hat{\xi}_{ij}$ as a function of the separation angle $\theta$ between two tomographic redshift bins $i$ and $j$:

$$\hat{\xi}^{(ij)}(\theta) = \frac{\sum_{ab} w_a w_b \sqrt{\epsilon_{a,b}^\text{obs} \epsilon_{a,b}^\text{obs} + \epsilon_{a,b}^\text{obs} \epsilon_{a,b}^\text{obs}} \Delta_{ab}^{(ij)}(\theta)}{\sum_{ab} w_a w_b (1 + \hat{m}_a)(1 + \hat{m}_b) \Delta_{ab}^{(ij)}(\theta)}. \quad (4)$$

---

6 These simulations are tailored to fit the KiDS imaging data and are based on the MICE2 simulation (Fosalba et al. 2015a,b; Crocce et al. 2015; Carretero et al. 2015; Hoffmann et al. 2015).
Table 2. The revised multiplicative shear bias $m_{\text{est}}$ for the KiDS-1000 sample compared to the original values $m_{\text{old}}$ of Asgari et al. (2021).

<table>
<thead>
<tr>
<th>Bin</th>
<th>Photo-$z$ range</th>
<th>$m_{\text{old}}$</th>
<th>$m_{\text{est}}$</th>
<th>$\sigma_{m}$</th>
<th>$\Delta z = z_{\text{old}} - z_{\text{est}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1 $&lt; z_{\text{old}}$ $&lt; 0.3$</td>
<td>$-0.009$</td>
<td>$-0.010$</td>
<td>0.019</td>
<td>0.000 $\pm$ 0.0096</td>
</tr>
<tr>
<td>2</td>
<td>0.3 $&lt; z_{\text{old}}$ $&lt; 0.5$</td>
<td>$-0.011$</td>
<td>$-0.009$</td>
<td>0.020</td>
<td>0.002 $\pm$ 0.0114</td>
</tr>
<tr>
<td>3</td>
<td>0.5 $&lt; z_{\text{old}}$ $&lt; 0.7$</td>
<td>$-0.015$</td>
<td>$-0.011$</td>
<td>0.017</td>
<td>0.013 $\pm$ 0.0116</td>
</tr>
<tr>
<td>4</td>
<td>0.7 $&lt; z_{\text{old}}$ $&lt; 0.9$</td>
<td>0.002</td>
<td>0.008</td>
<td>0.012</td>
<td>0.011 $\pm$ 0.0084</td>
</tr>
<tr>
<td>5</td>
<td>0.9 $&lt; z_{\text{old}}$ $&lt; 1.2$</td>
<td>0.007</td>
<td>0.012</td>
<td>0.010</td>
<td>$-0.006 \pm 0.0097$</td>
</tr>
</tbody>
</table>

Notes. The uncertainties $\sigma_{m}$ remain unchanged. The last column summarises the parameters of the correlated Gaussian redshift priors ($\mu, \pm \sigma$, where $\sigma$ is the square root of the diagonal element of the covariance matrix).

Here, $\Delta_{\text{est}}(\bar{z})$ is a function that expresses whether a pair of galaxies, $a$ and $b$, falls into an angular bin labelled by $\bar{z}$. Each galaxy has a weight $w$ and measured ellipticities, $\epsilon_{\text{est}}^a$ and $\epsilon_{\text{est}}^b$. The denominator applies the multiplicative shear bias $m$, which corrects the measured shear to match the true galaxy shear.

We extracted the cosmological information from the shear correlation signal using complete orthogonal sets of $E/B$-integrals (COSEBis, Schneider et al. 2010). These present a method to cleanly decompose the shear 2PCFs into $E$- and $B$-modes by applying a set of oscillatory filter functions defined over a finite angular range between $\theta_{\text{min}}$ and $\theta_{\text{max}}$. The filter functions $T_{\text{est}}(\theta)$ for the $n$-th COSEBI mode with

$$E_n = \frac{1}{2} \int_{\theta_{\text{min}}}^{\theta_{\text{max}}} \, \, d\theta \left[ T_{\text{est}}(\theta) \xi(\theta) + T_{-\text{est}}(\theta) \xi(-\theta) \right]$$

and

$$B_n = \frac{1}{2} \int_{\theta_{\text{min}}}^{\theta_{\text{max}}} \, \, d\theta \left[ T_{\text{est}}(\theta) \xi(\theta) - T_{-\text{est}}(\theta) \xi(-\theta) \right]$$

have exactly $n + 1$ roots.

One of the advantages of this formalism compared to the classical 2PCFs is that COSEBis are less sensitive to small scales, where the complex physics of baryon feedback plays an important role, if a subset of the modes is chosen accordingly (Asgari et al. 2020).

4.2. Analysis pipeline

Our analysis pipeline is an upgraded version of CosmoPipe$^8$ (Wright et al. 2020b) which is a wrapper for Cat_to_Obs$^9$ (Giblin et al. 2021) and the KiDS Cosmology Analysis Pipeline$^{10}$ (KCAP, Joachimi et al. 2021; Asgari et al. 2021; Heymans et al. 2021; Tröster et al. 2021) that have both been used previously to analyse the KiDS-1000 data. The pipeline measures the shear 2PCFs using TREECORR (Jarvis et al. 2004; Jarvis 2015) on angular scales between $0.5'$ and $300'$ from which we computed the first five COSEBI modes using the logarithmic versions of the filter functions $T_{\text{est}}(\theta)$. The logarithmic versions achieve a better compression of the cosmological signal onto fewer COSEBI modes.

We used the COSMOSIS framework (Zuntz et al. 2015) to compute theoretical predictions with the KCAP COSEBI module (Asgari et al. 2012). The linear matter power spectrum was modelled with CAMB (Code for Anisotropies in the Microwave Background, Lewis et al. 2000; Howlett et al. 2012) and its non-linear evolution with HMCODE (Mead et al. 2015, 2016), whereas intrinsic alignments were calculated based on the model of Hirata & Seljak (2004), Bridle & King (2007). We then compared these predictions to the measured COSEBis by sampling a Gaussian likelihood with MULTINEST (Feroz et al. 2009) using the analytical covariance model and priors of Joachimi et al. (2021). From this we inferred constraints on the cosmological parameters of a spatially flat CDM model. We additionally marginalised over a set of sample-dependent nuisance parameters which capture uncertainties in the shear and redshift calibration. Since Monte-Carlo samplers like MULTINEST are not designed to find the best fitting model parameters, we additionally ran a Nelder-Mead minimiser (Nelder & Mead 1965) starting from the maximum posterior point of all chains.

Based on this we quote parameter constraints and their uncertainty as the fit parameter value and the projected joint highest posterior density (PJ-HPD) that we obtained from the MULTINEST chains. It is important to note that both best fit parameters as well as the PJ-HPD have statistical uncertainties of about 0.1$\sigma$ or 10% on the $1\sigma$ constraints due to the limited number of posterior samples (Joachimi et al. 2021).

4.2.1. Redshift uncertainty

We propagated uncertainties in the redshift calibration to the cosmological constraints by allowing the redshift distribution of each tomographic bin $i$ to vary by a shift $\delta z_i$. We used a set of correlated Gaussian priors $\delta z_i \sim N(\mu_i, \sigma_i)$ which allowed us to apply an empirical redshift bias correction by choosing offsets $\mu_i \neq 0$. These offsets and their correlations (Table 2) were calibrated from spectroscopic and KiDS-like mock data (van den Busch et al. 2020) in Hildebrandt et al. (2021). Since our analysis uses different calibration samples with altered sample selections, we would in principle need to perform a similar mock data analysis to recalibrate the priors for $\delta z_i$. However, these new samples contain many new spectroscopic datasets and the inclusion of photometric redshifts presents an additional challenge when attempting to model realistic photo-$z$ errors. Therefore, we assumed that the variance of the KiDS-1000 priors is conservative enough to absorb potential changes of the redshift biases from KiDS-1000 to the new gold samples.

4.2.2. Multiplicative shear uncertainty

The second set of sample-dependent nuisance parameters is the average multiplicative shear bias ($m$-bias, see Eq. (4)) in each tomographic bin. The effect of the $m$-bias and its uncertainty on the COSEBis is captured in the covariance matrix and is calibrated by comparing the true galaxy ellipticities to those measured from a suite of image simulations generated by Kannawadi et al. (2019). The $m$-bias values vary little from sample to sample but are by up to 0.5$\sigma$ larger than those of the KiDS-1000 sample (Giblin et al. 2021). This led to the discovery of an issue with the way KiDS galaxies were assigned to galaxies in the COSMOS field which provided us with accurate shape information. We recomputed the $m$-bias (Table 2, Fig. 3) and find that the revised values are in good agreement with those of the new gold samples. The updated values are also well within the uncertainty on $m$ that was accounted for in A21.

---

7 In practice we did not apply the $m$-bias per galaxy in Eq. (4) but instead took the average value in each tomographic bin to avoid effects such as galaxy detection biases (Kannawadi et al. 2019).
8 https://github.com/AngusWright/CosmoPipe
9 https://github.com/KiDS-WL/Cat_to_Obs_K1000_P1
10 https://github.com/KiDS-WL/kcap
5. Results

In this section we present the new KiDS gold samples and cosmological constraints from an analysis of COSEBIs.

5.1. New KiDS gold samples

Similar to the calibration samples (Sect. 2.4) we divide the gold samples in two categories: primary, which are based on the full compilation of spectroscopic data plus optionally photo-z (spec-z fiducial, spec-z+PAUS, and spec-z+PAUS+COS15, see Sect. 2.4.1), and secondary samples, which are calibrated with subsets of the spectroscopic calibration sample (spec-z nQ ≥ 4) or by using exclusively photo-z (only-PAUS, only-COS15, and only-PAUS+COS15, see Sect. 2.4.2).

5.1.1. Primary gold samples

We made a quantitative comparison of the selection of the three primary KiDS gold samples based on the representation fraction of each tomographic bin, the effective sample number density compared to the density of the full KiDS-1000 source sample. These are summarised for all gold samples in Fig. 4. The numbers show that our new spectroscopic redshift calibration provides a much greater coverage of the KiDS source sample since our fiducial gold sample has a 9% higher accumulated number density than the previous data set calibrated by Hildebrandt et al. (2021), increasing the total representation fraction from 80% to 89%. In comparison to the former, our spec-z fiducial gold sample and those constructed by the addition of the PAUS and COSMOS2015 photo-z steadily increase the coverage fraction of KiDS galaxies across all tomographic bins, rising from 73%, 81%, and 8% to 88%, 90%, and 95% in bins two, three, and four respectively. The fifth tomographic bin, which contributes most of the cosmological signal in the cosmic shear analysis, shows the least change in its representation fraction due to the already very high coverage of 95% reported by Hildebrandt et al. (2021).

The only exception to the steadily increasing representation fractions is the first bin of spec-z+PAUS+COS15, where the number density is about 2.5% lower compared to the fiducial case. This is the result of the quality control (Eq. (2)) removing some SOM groupings due to discrepancies in ($z_{\text{cal}}$) and ($z_B$), which arise when adding calibration sources and/or changing the gold selection. Expanding the calibration sample may shift ($z_{\text{cal}}$) significantly, in particular in sparsely occupied SOM groupings, such that ($z_{\text{cal}}$)−($z_B$) exceeds the quality control threshold 5$\sigma_{\text{mad}}$, which will flag and exclude the corresponding KiDS galaxies from the gold sample.

The SOM redshift distributions of the gold samples are shown in Fig. 5; it is important to note that these samples do not represent the same galaxies. A comparison reveals two effects when adding photo-z to the calibration sample: First, the bulk of the redshift distributions is skewed to lower redshifts as we added more data to the calibration sample which is most evident in the third tomographic bin. Secondly, COSMOS2015 added a significant portion of high redshift objects to the compilation that extends the coverage of KiDS galaxies to higher redshifts, enhancing the tails of the redshift distributions and significantly increasing the mean redshifts, in particular in bin five.

Finally we compared the calibration sample redshift distributions in each tomographic bin to the resulting gold sample redshift distributions (Fig. 6). For most of the samples these distributions have very similar shapes and with mean redshifts agreeing within ±0.02, which indicates that the re-weighting (Eq. (1)) of the SOM groupings is very small on average. This changes once the COSMOS2015 redshifts are added to the calibration sample. Due to their significantly higher depth and mean redshifts, the redshift tails must be down-weighted significantly (up-weighted in bin one and two) to match the density of the KiDS source sample. The down-weighting of the low redshift tails in the upper three bins can be explained by the fact that COSMOS2015 adds faint galaxies at these redshifts. The corresponding KiDS galaxies have a lower lensing weight, which must be compensated by the SOM cell weights.

5.1.2. Secondary gold samples

The first of our secondary gold samples we calibrated using only those galaxies of the spectroscopic compilation that have the most secure spectroscopic redshifts of at least 99% confidence. Due to similarities to the SOM calibration sample used by Hildebrandt et al. (2021), see Sect. 2.4.3, this spec-z nQ ≥ 4 gold sample positions itself in between the latter and the spec-z fiducial sample in terms of representation fractions. In tomographic bin four and five however it is lacking some of the high redshift sources due to the more conservative spectroscopic calibration sample and those constructed by the addition of the PAUS and COSMOS2015 photo-z, which extends the coverage of KiDS galaxies to higher redshifts, enhancing the tails of the redshift distributions and significantly increasing the mean redshifts, in particular of bin five.

The remaining three gold samples exclusively rely on photometric redshifts from PAUS and COSMOS2015 (Sect. 2.4.2). Due to the great overlap of sources, both samples that contain the
Fig. 5. Comparison of the gold sample redshift distributions and their mean redshifts for all tomographic bins obtained from the different subsets of the calibration sample. All gold samples represent a different subset of the full KiDS-1000 source sample.

Fig. 6. Comparison of the tomographically binned calibration sample (black lines) to the gold sample redshift distributions (coloured lines). The greater the difference between the black and the coloured lines, the more weighting is applied by the SOM to match the calibration sample to the KiDS data.
Table 3. Summary of the main cosmological parameter constraints (best fit and 68th-percentile PJ-HPD) from COSEBIs for all gold samples and their comparison to Asgari et al. (2021) and Planck legacy (TT, TE, EE + lowE).

<table>
<thead>
<tr>
<th>Sample</th>
<th>$\chi^2$</th>
<th>$A_{1A}$</th>
<th>$\Omega_m$</th>
<th>$\sigma_8$</th>
<th>$S_8$</th>
<th>$\Sigma_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>spec-z fiducial</td>
<td>63.2</td>
<td>0.301$^{+0.343}_{-0.377}$</td>
<td>0.268$^{+0.126}_{-0.055}$</td>
<td>0.791$^{+0.110}_{-0.155}$</td>
<td>0.748$^{+0.021}_{-0.025}$</td>
<td>0.744$^{+0.017}_{-0.021}$</td>
</tr>
<tr>
<td>spec-z+PAUS</td>
<td>64.0</td>
<td>$-0.097^{+0.422}_{-0.416}$</td>
<td>0.226$^{+0.108}_{-0.053}$</td>
<td>0.857$^{+0.130}_{-0.164}$</td>
<td>0.743$^{+0.031}_{-0.016}$</td>
<td>0.732$^{+0.028}_{-0.012}$</td>
</tr>
<tr>
<td>spec-z+PAUS+COS15</td>
<td>68.3</td>
<td>0.164$^{+0.450}_{-0.428}$</td>
<td>0.194$^{+0.109}_{-0.048}$</td>
<td>0.940$^{+0.144}_{-0.200}$</td>
<td>0.757$^{+0.016}_{-0.026}$</td>
<td>0.740$^{+0.013}_{-0.025}$</td>
</tr>
<tr>
<td>spec-z nQ ≥ 4</td>
<td>61.6</td>
<td>$-0.033^{+0.408}_{-0.391}$</td>
<td>0.195$^{+0.079}_{-0.046}$</td>
<td>0.951$^{+0.113}_{-0.171}$</td>
<td>0.766$^{+0.023}_{-0.023}$</td>
<td>0.750$^{+0.021}_{-0.019}$</td>
</tr>
<tr>
<td>Only PAUS</td>
<td>74.1</td>
<td>$-0.186^{+0.526}_{-0.345}$</td>
<td>0.203$^{+0.108}_{-0.053}$</td>
<td>0.916$^{+0.163}_{-0.183}$</td>
<td>0.752$^{+0.028}_{-0.019}$</td>
<td>0.737$^{+0.023}_{-0.017}$</td>
</tr>
<tr>
<td>Only COS15</td>
<td>67.2</td>
<td>0.164$^{+0.539}_{-0.545}$</td>
<td>0.253$^{+0.132}_{-0.074}$</td>
<td>0.814$^{+0.157}_{-0.177}$</td>
<td>0.747$^{+0.033}_{-0.033}$</td>
<td>0.741$^{+0.027}_{-0.023}$</td>
</tr>
<tr>
<td>Only PAUS+COS15</td>
<td>67.1</td>
<td>0.076$^{+0.515}_{-0.369}$</td>
<td>0.216$^{+0.176}_{-0.033}$</td>
<td>0.876$^{+0.085}_{-0.243}$</td>
<td>0.744$^{+0.020}_{-0.023}$</td>
<td>0.731$^{+0.024}_{-0.017}$</td>
</tr>
<tr>
<td>Asgari et al. (2021)</td>
<td>82.2</td>
<td>0.264$^{+0.424}_{-0.336}$</td>
<td>0.246$^{+0.101}_{-0.060}$</td>
<td>0.838$^{+0.140}_{-0.141}$</td>
<td>0.759$^{+0.024}_{-0.021}$</td>
<td>0.751$^{+0.024}_{-0.016}$</td>
</tr>
<tr>
<td>Planck legacy</td>
<td>–</td>
<td>–</td>
<td>0.315$^{+0.006}_{-0.010}$</td>
<td>0.813$^{+0.007}_{-0.008}$</td>
<td>0.838$^{+0.013}_{-0.020}$</td>
<td>0.841$^{+0.018}_{-0.021}$</td>
</tr>
</tbody>
</table>

Notes. Additionally shown are the $\chi^2$ values (for 70.5 effective degrees of freedom) and $\Sigma_8 = \sigma_8(\Omega_m/0.3)^{0.5}$ calculated for $\sigma_8 = 0.55$, determined from a fit to the posterior samples of the fiducial chain.

Fig. 7. Marginalised constraints for the joint distributions of $S_8$ and $\Omega_m$ (68% and 95% credible regions) obtained for different gold samples and Planck legacy (TT, TE, EE + lowE). As the contours represent different galaxy samples, some deviation is expected.

Fig. 8. Comparison of the $S_8$ constraints from our gold samples to other studies. We show the best fit (where available) and 68th-percentile PJ-HPD (circles, opaque data points), and the maximum of the marginal distribution and the associated 68th-percentile (diamonds, semi-transparent). We compare to Asgari et al. (2021), HSC-Y1 (Hikage et al. 2019), DES-Y3 (Amon et al. 2021), and Planck legacy. The coloured vertical line and outer bands indicate the constraints from the fiducial gold sample, the inner bands the expected variance of the sampler.

5.2. Cosmological constraints

We present cosmological results for our primary KiDS gold samples, focusing on a relative comparison to A21 and other literature values. We summarise the numerical values of the most relevant cosmological parameters in Table 3. In Figs. 7 and 8 we highlight comparisons of the derived parameter $S_8 = \sigma_8(\Omega_m/0.3)^{0.5}$, which is the primary measurable of weak lensing due to the degeneracy between $\Omega_m$, the dimensionless matter density parameter, and $\sigma_8$, parameterising the amplitude of the linear power spectrum.

5.2.1. Reanalysis of the KiDS-1000 data

First we ensured that our wrapper for the cosmological pipeline delivers results that are consistent with those of A21. We reanalysed the original KiDS-1000 data, running our pipeline with the same input parameters (‘Asgari reanalysis’). We find that the constraints on all cosmological parameters agree within the expected variance of the Monte-Carlo sampler. The best fit solution has a slightly larger $\chi^2$ of 83.8 compared to 82.2 for KiDS-1000 which is driven primarily by the slight change in the data vector (see Appendix C). Similarly, the correction of the $m$-bias values (Table 2) has no significant impact on the cosmological constraints either (Table C.1).
5.2.2. Primary gold samples

For our KiDS gold samples we find a tendency to lower $S_8$-values, albeit at low significance (Figs. 7, 8 and Table 3). In particular we obtained $S_8 = 0.748^{+0.025}_{-0.023}$ for the spec-$z$ fiducial and $S_8 = 0.743^{+0.007}_{-0.006}$ for the spec-$z$+PAUS gold samples, which is about $0.4\sigma$ lower compared to A21, and exceeds the expected statistical variance of the sampler (about $0.1\sigma$). Deviations to some degree between these gold samples are expected, since they all represent different galaxy selections. The goodness of fit improves for all gold samples but in particular for spectral-$z$ fiducial, where it reduces to $\chi^2 = 63$ compared to 82 in KiDS-1000. The spec-$z$+PAUS+COS15 sample on the other hand is in very good agreement with the original KiDS-1000 analysis.

In addition to $S_8$ we considered the more general $\Omega_m - \sigma_8$-degeneracy case of $\Sigma_8 = \sigma_8(\Omega_m/0.3)^{2}$, where $\sigma_8$ is a free parameter. We estimated the reference value for $\alpha$ by fitting the posterior samples of the fiducial chain and obtain $\alpha_{\text{fid}} = 0.55$ which is close to $\alpha = 0.54$ for COSEBIs in A21. This projection optimises the signal-to-noise ratio compared to $S_8$ and we find $\Sigma_8 = 0.744^{+0.017}_{-0.021}$ for spec-$z$ fiducial compared to $\Sigma_8 = 0.751^{+0.024}_{-0.016}$ which we obtained for A21 with $\alpha = \alpha_{\text{fid}}$. Furthermore, the scatter of $\Sigma_8$ for the different gold samples is significantly smaller than the scatter in $S_8$ (see Table 3).

If we compare the marginal errors of $S_8$ for the different gold samples (Table D.1), which have a smaller statistical variance than the PJ-HPD, we find that the constraints improve by 5% when including PAUS and another 4% when including COSMOS2015. Since the constraints on $\Sigma_8$ are almost constant, these changes in $S_8$ are most probably related to small changes in the $\Omega_m - \sigma_8$-degeneracy. We also compared $\alpha_{i}\Sigma_8$, which is the dimensionless amplitude of the intrinsic alignment galaxy power spectrum, and find that its value is stable within the uncertainties in all our analyses.

6. Discussion

6.1. Gold sample selection and calibration

A side-by-side comparison of the different KiDS gold samples presented in Sect. 5.1 is non-trivial. On the one hand adding or removing galaxies from the calibration sample changes the redshift distribution of each of the SOM groupings that, according to Eq. (1), determine the sample’s redshift distribution. On the other hand two distinct gold samples are comprised of different galaxies since a modification of the calibration sample will also apply an implicit selection on the set of representative SOM groupings. Both of these effects combined determine the overall calibrated redshift distribution.

This is exemplified by the fact that, as we expand the calibration sample from the spec-$z$ nQ $\geq$ 4 subset to spec-$z$+PAUS+COS15, we generally see that both the representation fractions and the mean redshifts increase across all bins. The galaxies added in each iteration are typically fainter (with the exception of PAUS) at the cost of lower redshift accuracy, which allowed us to calibrate additional KiDS galaxies, preferentially at the tails of the redshift distributions. At the same time we can observe in Fig. 5 that the redshift distributions are skewed to lower redshifts, which can be explained by the fact that there are disproportionately many more galaxies with $z < 0.5$ added in each iteration to the calibration sample (Fig. 2). These in turn increase the representation fraction of low and intermediate redshift galaxies in the gold sample (see also Fig. 4). This implies that we are changing the redshift calibration since the skewing applies to each individual SOM group. We were able to separate these two effects by splitting the spec-$z$ fiducial and spec-$z$+PAUS+COS15 gold samples in two subsets, one containing those SOM groupings that are common to both samples (i.e. containing the same KiDS galaxies) and groupings that can only be calibrated using the full redshift compilation. The subset of KiDS galaxies that is common to both gold samples shows the same redshift skewing as seen with all galaxies (top panels of Fig. 9), whereas the additional COSMOS2015 galaxies contribute significantly at the low and high redshift tails of the tomographic bins (bottom panels of Fig. 9).

The final ingredient to the redshift calibration is the quality control cut (Eq. (2)) that we applied to remove potentially mis-calibrated parts of the colour space. This becomes most obvious when comparing the representation fractions of the first tomographic bin (Fig. 4) which decreases whenever we added the COSMOS2015 data (compare spec-$z$+PAUS to spec-$z$+PAUS+COS15 and only-PAUS to only-PAUS+COS15). The low redshift of the KiDS galaxies in this bin makes the sample particularly susceptible to the addition of high redshift galaxies. These can significantly change the mean redshift of the calibration sample $(\langle z_{\text{cal}} \rangle)$ compared to the KiDS photo-z in the SOM groupings which then may fail to pass the quality control (see also Appendix B). The great depth of the COSMOS2015 data compared to the spectroscopic data also explains why the SOM needs to apply more weighting to match the spec-$z$+PAUS+COS15 compilation to the KiDS colour-space (Fig. 6).

6.2. Cosmological constraints

The gold sample selection effects are, due to their redshift dependence, directly propagated to the cosmological constraints (Sect. 5.2), causing shifts in $S_8$ of up to $0.5\sigma$ from sample to sample. One of the assumptions in our analysis is that we can adopt the same Gaussian priors for the $\delta_z\alpha$ nuisance parameters (Sect. 4.2) that are used by A21. We therefore reanalysed the fiducial gold sample assuming no knowledge of the empirical redshift bias by centring the priors on $\mu_z = 0$. For this run we find that the value of $S_8 = 0.747^{+0.025}_{-0.023}$ is in good agreement with the fiducial analysis.

On the other hand, it may seem that our choice for the widths $\sigma_8$ of the $\delta_z\alpha$ priors may be insufficient to accommodate for the apparent variance in the mean redshifts of the different gold samples (see Fig. 5). This variance, however, is not only determined by potential systematic biases in the redshift calibration between any of these samples, but also by changes in the gold sample selection itself, as discussed above. Therefore, the question of the correct redshift prior can only be answered with realistic simulated data sets that are currently not available for our extended redshift calibration sample. Nevertheless, a comparison of the $S_8$ values allows us to get an estimate of the variance induced by the selection effects in the calibration data and the resulting parameter constraints from the gold samples.

6.2.1. The fiducial gold sample

Next to the spec-$z$ fiducial sample that is based on the full spectroscopic compilation we also defined the spec-$z$ nQ $\geq$ 4 sample that relies only on the most secure redshifts. The estimated $S_8 = 0.766^{+0.025}_{-0.023}$ for the latter, arising from particularly low $\Omega_m$ and high $\sigma_8$ values, is about $0.5\sigma$ larger than in the fiducial case and surpasses all other gold samples (Table 3). In the $\Sigma_8$ projection this difference reduces to $0.2\sigma$.
The shift in $S_8$ between gold samples that are both calibrated with spectroscopic data begs the question which of these estimates is more reliable. The primary difference between the two calibration datasets is the selection using redshift quality flags. Selecting spectra based on the redshift confidence is a trade-off between constructing a sample that is confined to regions of the colour-redshift-space in which galaxies have distinct spectral features that allow secure redshift determination and a sample with an increasing fraction of galaxies with catastrophically misidentified redshifts. In the latter case the redshifts of the calibration sample themselves cause a biasing of the gold sample redshifts and in turn $S_8$. In case of selecting only the highest quality redshifts, the biases arises from a misrepresentation of the imaging data by the calibration sample, as shown by Hartley et al. (2020). The redshift distribution of the calibration sample in each SOM grouping depends on the quality flag and thus the relative representation of different galaxy populations in the gold sample may change. We investigated the magnitude of both these effects by assuming two worst-case scenarios which shift down the $S_8$ estimate obtained for spec-$z$ fiducial and shift up $S_8$ for spec-$z$ nQ $\geq 4$.

Scenario A assumes in the case of the spec-$z$ fiducial sample that 5% of truly low redshift galaxies with $3 \leq n_Q < 4$ (nominal 95% certainty), as well as 1% in case the of nQ $\geq 4$ (99% certainty), are catastrophically misidentified as high redshift galaxies. Since both redshift flags are equally common in the fiducial sample we expect a combined spectroscopic failure rate of about 3%. We implemented this worst-case scenario on spec-$z$ fiducial by truncating the top 3% of all redshift distributions which should increase the recovered $S_8$ value. We calculated the redshift $z_{97}$ corresponding to the 97-th percentile of the redshift distribution $n(z)$, set $n(z) = 0$ at $z > z_{97}$, and re-normalised to reproduce the original gold sample number density.

In scenario B we speculate that the calibration of the spec-$z$ nQ $\geq 4$ sample suffers from the same spectroscopic misrepresentation effects studied by Hartley et al. (2020, from Fig. 6 therein), who found redshift biases $\langle z \rangle - \langle z_{true} \rangle$ of 0.008, 0.022, and -0.003, and -0.058 in the four tomographic redshifts bins of simulated DES and spectroscopic data, for the first time implementing a realistic, simulated nQ $\geq 4$ sample selection. Since we currently do not have comparable spectroscopic mock data in KiDS, we assumed in this scenario that the bias applies at the same magnitude to the spec-$z$ nQ $\geq 4$ gold sample. We therefore corrected the assumed bias by interpolating the values from the four DES bins to the five tomographic bins of KiDS and shift the spec-$z$ nQ $\geq 4$ redshift distributions by $-0.008$, $-0.015$, $-0.014$, 0.003, and 0.058. We considered this to be an even more conservative assumption than scenario A, since the bias should be significantly smaller in case of KiDS thanks to the nine-band imaging and the improvements of the SOM calibration over the classical DIR approach (Sect. 3.1) that is used by Hartley et al. (2020).
In our analysis the shifts in $\Sigma$ finally, we made a relative comparison of the $\Sigma$ cosmological analysis. carefully to minimise biases in the redshift calibration and the out colour-redshift degeneracies in each SOM grouping, and the significant impact on cosmological constraints. Furthermore that the selection ejection $\Sigma$ spec- 

2020) show a more pronounced shift in the opposite direction is caused by the gold sample selection $\Sigma$ fiducial (Fig. 11). While the cosmological constraints in a similar comparison between redshifts calibrated using spectroscopic data and COSMOS2015 (Hildebrandt et al. 2020) show a more pronounced shift in the opposite direction, this difference is caused by the gold sample selection (Wright et al. 2020b). In our analysis the shifts in $\Sigma$ may be explained by the fact that COSMOS2015 tends to calibrate the KiDS galaxies to higher redshifts than the spectroscopic calibration data alone (see Fig. 9), translating to lower $\Sigma_8$. The same reasoning does not explain why $\Sigma_8$ reduces further when the PAUS data are included, therefore this behaviour is most likely owed to the quality control (Eq. (2)). When the photo-z data are combined with the spectroscopic compilation, it may result in a significantly different distribution of $|\langle z_{\text{cal}} \rangle - \langle z_{\text{bf}} \rangle|$ for the SOM groupings, which has non-trivial implications for the gold sample selection and the derived cosmological constraints.

With these modifications to the redshift distributions, scenario A (fiducial, truncated top 3%) yields a higher and scenario B (nQ ≥ 4, Hartley corrected) a lower estimate for $\Sigma_8$ (Fig. 10, left). These results indicate that the combination of these two effects may explain the observed differences between $\Sigma_8$ in spec-z fiducial and spec-z nQ ≥ 4. When comparing the projection $\Sigma_8$ instead (Fig. 10, right), which is less susceptible to shifts along the $\Omega_m$-$\sigma_8$-degeneracy, the difference between spec-z fiducial and spec-z nQ ≥ 4 is much smaller. However the shift, introduced when correcting the redshift distributions in scenario B, is about twice as big compared to the $\Sigma_8$ case, indicating that the selection effects studied in Hartley et al. (2020) can have a significant impact on cosmological constraints. Furthermore the completeness, which determines the ability to correctly map out colour-redshift degeneracies in each SOM grouping, and the quality of the calibration sample redshifts should be balanced carefully to minimise biases in the redshift calibration and the cosmological analysis.

6.2.2. Other gold samples

Finally, we made a relative comparison of the $\Sigma_8$ constraints from the remaining gold samples, since $\Sigma_8$ typically exhibits a smaller scatter than the corresponding $\Sigma_8$ values. The gold samples that are calibrated using only photo-z from PAUS, COSMOS2015, or a combination of both, prefer smaller $\Sigma_8$ values compared to spec-z fiducial (Fig. 11). While the cosmological constraints in a similar comparison between redshifts calibrated using spectroscopic data and COSMOS2015 (Hildebrandt et al. 2020) show a more pronounced shift in the opposite direction, this difference is caused by the gold sample selection (Wright et al. 2020b). In our analysis the shifts in $\Sigma_8$ may be explained by the fact that COSMOS2015 tends to calibrate the KiDS galaxies to higher redshifts than the spectroscopic calibration data alone (see Fig. 9), translating to lower $\Sigma_8$. The same reasoning does not explain why $\Sigma_8$ reduces further when the PAUS data are included, therefore this behaviour is most likely owed to the quality control (Eq. (2)). When the photo-z data are combined with the spectroscopic compilation, it may result in a significantly different distribution of $|\langle z_{\text{cal}} \rangle - \langle z_{\text{bf}} \rangle|$ for the SOM groupings, which has non-trivial implications for the gold sample selection and the derived cosmological constraints.

7. Conclusion

We applied the SOM redshift calibration technique (Wright et al. 2020a) to define and calibrate the redshift distributions of a new set of KiDS-1000 gold samples by adopting a new spectroscopic calibration sample. Compared to previous work by Hildebrandt et al. (2021) we doubled the size of this calibration sample by adding more than ten additional spectroscopic campaigns such as C3R2 and DEVILS, which allowed us to calibrate an additional 9% of the KiDS galaxies.

We took this one step further by enhancing the calibration sample with precision photometric redshifts from the PAU survey and COSMOS2015, maintaining a hierarchy that prefers spectroscopic over PAUS and COSMOS2015 redshifts to resolve duplicates in the three catalogues. The resulting KiDS gold sample increases by additional 6% and covers nearly 98% of all KiDS galaxies in the fifth tomographic bin.

When comparing these gold samples we find changes in the mean redshifts of up to $|\Delta z| = 0.026$ which originate from selection effects in the calibration sample. First, there are residual modifications to the redshift calibration of those KiDS sources that are found in both gold samples. These modifications are a direct consequence of changing the redshift distribution of the calibration sample when adding the photo-z. Second and most important, the selection of KiDS sources itself changes since the faint COSMOS2015 data allows us to calibrate additional galaxies at both low and high redshifts. These results highlight the importance of quantifying and calibrating potential method dependent redshift biases arising from selection effects, as has been shown in previous work. This requires sophisticated galaxy mock data with sufficient redshift coverage, realistic galaxy colours and accurate modelling of photometric and spectroscopic galaxy samples, in particular if one aims to study the impact of photo-z outliers on the calibration sample.

In the second part of this study we performed a cosmic shear analysis using COSEBIs and find $S_8 = 0.748^{+0.021}_{-0.022}$ for our fiducial and $v$ for the photo-z-enhanced gold sample which is slightly lower than, but still in excellent agreement with, previous work on KiDS-1000 by Asgari et al. (2021). As part of our additional systematic testing we created a third gold sample that we calibrated using only the most secure spectra of our spectroscopic compilation (nQ ≥ 4). We measure $S_8 = 0.766^{+0.023}_{-0.022}$ which corresponds to an increment of 0.5σ compared to the fiducial
case. Our analysis speculates that this sample underestimates the true redshifts and overestimates $S_8$ due to an implicit selection introduced when limiting the calibration to the most secure spectroscopic redshifts (Hartley et al. 2020). Therefore completeness and quality of the calibration sample can be optimised, this however requires realistic simulations to quantify the impact of this trade-off on the recovered redshift distributions.

Finally we analysed four KiDS gold samples which are all calibrated from different subsets of our extended calibration sample or using only photo-$z$ from PAUS and COSMOS2015. No matter how we calibrated the KiDS source galaxies, we find that all seven gold samples studied in this work scatter in the range of $S_8 = 0.743 \ldots 0.766$ around our fiducial analysis. This further confirms previously reported tensions of KiDS with measurements of the CMB by Planck Collaboration VI (2020) at 3.0$\sigma$ to 3.6$\sigma$ for a flat $\Lambda$CDM model.

In summary, there seems to be little benefit in using precision photometric redshifts for the SOM redshift calibration given the error bars on our spectroscopic redshifts alone. The spec-$z$+PAUS+Cosmos15 gold sample achieves, compared to our fiducial sample, a 6% improvement in terms of the number density, yet improvements on the cosmological constraints are marginal. Nevertheless, if the spectroscopic coverage is significantly lower or one wishes to target higher redshifts, photo-$z$ samples are a valuable source of complementary calibration data. However, the greater the dependence on photometric redshifts is, the more attention must be paid to redshift outliers to guarantee a good balance between statistical uncertainties and systematic biases in the redshift calibration. Given their challenging calibration requirements this is in particular true for the next generation, stage IV surveys such as Euclid (Laureijs et al. 2011) or the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST, Ivezić et al. 2019).

Despite photo-$z$ outliers, photometric redshift samples have one significant advantage over spectroscopically selected samples: they achieve a much higher completeness, which can mitigate selection effects in the calibration sample and may improve the ability of the SOM to map out the full extent of colour-redshift degeneracies.

**Acknowledgements.** We acknowledge support from the European Research Council under grant numbers 770935 (JvdB, AW, HH) and 647112 (MA), and the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT) introduced when limiting the calibration to the most secure spectroscopic redshifts (Hartley et al. 2020).

**References**


Edge, A., Sutherland, W., Kuijken, K., et al. 2013, The Messenger, 154, 32


Appendix A: Updated spectroscopic compilation

Here we present the details of our extended and revised spectroscopic compilation discussed in Sect. 2.2. It includes redshifts from surveys covering a number of deep extragalactic fields for which we have nine-band VST+VISTA photometry of similar quality as the main KiDS+VIKING data (KiDZ). In this paper we used six such fields shown in Fig. A.1: three of them (COSMOS, CDF-S and VVDS-2h) are covered by a number of partly overlapping surveys and we merged them into our redshift calibration sample. In doing this, we translated the input redshift quality flags or assessments into our flag NQ, with NQ ≥ 4 indicating the most secure spectroscopic redshifts (confidence 99% or more), while 3 ≤ NQ < 4 indicate secure, but lower-confidence redshifts. The particular surveys are listed in Table A.1, in the ascending order of their mean redshift determined after removing multiples and cross-matching with our imaging. The details of the selection and quality assignment for the particular samples are as follows:

- hCOSMOS (Damjanov et al. 2018): we used all the galaxies from the published dataset and assigned them redshift quality NQ = 4;
- GAMA-G15Deep (Kafle et al. 2018; Driver et al. 2022): we selected galaxies with input redshift quality Z_QUAL ≥ 3 and with redshifts z > 0.001 to avoid stellar contamination. We assigned NQ = Z_QUAL;
- G10-COSMOS (Davies et al. 2015): we used Z_BEST as the redshift value and followed the recommended selection for galaxy spectroscopic redshifts, that is Z_BEST > 0.0001, Z_USE < 3, and STAR_GALAXY_CLASS = 0. As the redshifts provided in that compilation do not always have a quality flag that could be easily translated into our NQ, we assigned all these galaxies a specific flag NQ = 3.5;
- ACES (Cooper et al. 2012): we selected redshifts with Z_QUALITY ≥ 3 and with redshift errors below 1%. We assigned NQ = Z_QUALITY;
- OzDSS (Lidman et al. 2020): we used two patches partly overlapping with KiDZ imaging (around CDF-S and around VVDS-2h). We selected redshifts z > 0.002 (stellar contamination removal) and required quality qop ∈ [3, 4]. We assigned NQ = qop;
- DEVILS (Davies et al. 2018): we selected spectroscopic redshifts only (zBestType = spec) and require the flags starFlag = 0, mask = 0, and artefactFlag = 0. As the DEVILS sample is a compilation including both own redshifts and external ones, we assigned NQ = 4 if zBestSource = DEVILS (redshift obtained from DEVILS observations) and NQ = 3 (external redshifts);
- VIPERS (Scoccimarro et al. 2018): we required sources with the flag 2 ≤ zflag < 10 or 22 ≤ zflag < 30. We assigned NQ = 4 if 3 ≤ zflag < 5 or 23 ≤ zflag < 25, and NQ = 3 otherwise;
- VVDS (Le Fèvre et al. 2005, 2013): we joined the WIDE, DEEP, and UDEEP sub-samples and used only sources with ZFLAGS ∈ {3, 4, 23, 24}. We assigned NQ = 4 to these redshifts;
- LEGA-C (van der Wel et al. 2016): we selected sources with the flag f_use = 1 and assigned NQ = 4 to them;
- DEEP2 (Newman et al. 2013): as in the previous KiDS papers (Hildebrandt et al. 2017, 2020, 2021), we used sources from two equatorial fields (0226 & 2330), required sources with ZQUALITY ≥ 3 and with redshift errors smaller than 1%. We assigned NQ = Z_QUALITY;
- C3R2: we combined four public datasets, DR1 (Masters et al. 2017), DR2 (Masters et al. 2019), DR3 (Stanford et al. 2021), and KMOS (Euclid Collaboration 2020) and excluded the Northern EGS field. We took sources with QFLAG ≥ 3 and assigned NQ = QFLAG;
- DEIMOS (Hasinger et al. 2018): we required the quality flag Q = 2, while for assigning NQ we used the other flag provided, namely NQ = 4 for Qe ∈ [4, 14], NQ = 3 otherwise;
- GOODS/CDF-S: we used a public ESO compilation of spectroscopy in the CDF-S field11 (Popesso et al. 2009; Balestra et al. 2010). From the compilation, we chose secure redshifts (assigning NQ = 4 to them) and likely redshifts (NQ = 3) following the recommendations in the dataset description12;
- FMOS-COSMOS (Silverman et al. 2015): we selected sources with the quality flag q_z ≥ 2 and assigned NQ = 4 if q_z = 4 or NQ = 3 otherwise;
- zCOSMOS: we used this name for a proprietary compilation of various spectroscopic surveys in the COSMOS field, kindly provided to us by Mara Salvato, updated as of 01 September 2017. That dataset includes some of the surveys already included in our compilation, but it also provides redshifts from various other campaigns. We used the provided quality flag and select sources meeting the criteria 3 ≤ Q_f ≤ 5 or 13 ≤ Q_f ≤ 15 or 23 ≤ Q_f ≤ 25 or Q_f ∈ {6, 10}, do not use sources from lower-confidence determinations (e.g. grism), and limit the redshifts to z spec > 0.002 to avoid stellar contamination. We assigned redshift quality as NQ = min((Qf mod 10), 4);
- VUDS (Le Fèvre et al. 2015): we used sources with redshift flag zflags ending with [3, 4, 9] (reliability ≥ 80%) and assigned NQ = 4 if 3 ≤ zflags ≤ 8 or 13 ≤ zflags ≤ 25, otherwise NQ = 3;
- VANDELS (Garilli et al. 2021): we selected sources for which (zflag mod 10) ∈ {2, 3, 4} and assigned NQ = 4 if (zflag mod 10) ∈ {3, 4} or NQ = 3 otherwise. The reassessment of the quality flags was motivated by the reportedly high redshift confidence of objects with flag values of two and three.

When joining the above samples into one dataset, we removed duplicates both from overlapping surveys as well as within the individual ones. In the former case, if for a given source there are redshifts from different surveys available, we assigned the most reliable measurement based on a specific ‘hierarchy’. Namely, we joined the catalogues by cross-matching objects within 1′ radius and apply the following order of preference:

- COSMOS field: G10-COSMOS > DEIMOS > hCOSMOS > VVDS > Lega-C > FMOS > VUDS > C3R2 > DEVILS > zCOSMOS;
- CDF-S field: ACES > VANDELS > VVDS > GOODS/CDF-S > DEVILS > OzDSS;
- VVDS-2h field: VIPERS > VVDS > C3R2 > DEVILS > OzDSS.

For objects with multiple spectroscopic measurements within a particular survey, we either took the redshift with the highest quality flag or, if various entries for the same source have the same quality flag and the reported redshifts differ by no more than 0.005, we calculated the average. If the reported redshifts have the same quality flag but differ by more than 0.005, we excluded such a source from the compilation.

11 Available from https://www.eso.org/sci/activities/garching/projects/goods/MasterSpectroscopy.html
12 https://www.eso.org/sci/activities/garching/projects/goods/MASTERCAT_v3.0.dat
Table A.1. Listing of the spectroscopic samples that enter the spectroscopic compilation summarising the sample sizes and mean redshifts. The values are calculated after removing duplicates between overlapping catalogues and matching with the KiDS imaging.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Count</th>
<th>(\langle z \rangle)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>hCOSMOS</td>
<td>503</td>
<td>0.308</td>
<td>Damjanov et al. (2018)</td>
</tr>
<tr>
<td>GAMA-G15Deep</td>
<td>1840</td>
<td>0.357</td>
<td>Kafle et al. (2018), Driver et al. (2022)</td>
</tr>
<tr>
<td>G10-COSMOS</td>
<td>14849</td>
<td>0.586</td>
<td>Davies et al. (2015)</td>
</tr>
<tr>
<td>ACES</td>
<td>4233</td>
<td>0.593</td>
<td>Cooper et al. (2012)</td>
</tr>
<tr>
<td>OzDES</td>
<td>930</td>
<td>0.638</td>
<td>Lidman et al. (2020)</td>
</tr>
<tr>
<td>DEVILS</td>
<td>5222</td>
<td>0.682</td>
<td>Davies et al. (2018)</td>
</tr>
<tr>
<td>VIPERS</td>
<td>2346</td>
<td>0.718</td>
<td>Scodéggio et al. (2018)</td>
</tr>
<tr>
<td>VVDS</td>
<td>5190</td>
<td>0.737</td>
<td>Le Fèvre et al. (2005, 2013)</td>
</tr>
<tr>
<td>LEGA-C</td>
<td>216</td>
<td>0.818</td>
<td>van der Wel et al. (2016)</td>
</tr>
<tr>
<td>DEEP2</td>
<td>8564</td>
<td>0.962</td>
<td>Newman et al. (2013)</td>
</tr>
<tr>
<td>C3R2</td>
<td>2512</td>
<td>0.980</td>
<td>Masters et al. (2017, 2019), Euclid Collaboration (2020), Stanford et al. (2021)</td>
</tr>
<tr>
<td>DEIMOS</td>
<td>1729</td>
<td>1.045</td>
<td>Hasinger et al. (2018)</td>
</tr>
<tr>
<td>GOODS</td>
<td>1999</td>
<td>1.292</td>
<td>ESO compilation of GOODS/CDF-S spectroscopy</td>
</tr>
<tr>
<td>FMOS-COSMOS</td>
<td>272</td>
<td>1.566</td>
<td>Silverman et al. (2015)</td>
</tr>
<tr>
<td>zCOSMOS</td>
<td>1875</td>
<td>1.602</td>
<td>private comm. from M. Salvato</td>
</tr>
<tr>
<td>VUDS</td>
<td>205</td>
<td>1.974</td>
<td>Le Fèvre et al. (2015)</td>
</tr>
<tr>
<td>VANDELS</td>
<td>336</td>
<td>2.504</td>
<td>Garilli et al. (2021)</td>
</tr>
</tbody>
</table>

Fig. A.1. On-sky distribution of spectroscopic sources matched against the six KiDZ fields.

Fig. A.2. Redshift distributions of the spectroscopic samples that enter the spectroscopic compilation, ordered by mean redshift (Table A.1). The distributions are calculated after removing duplicates between overlapping catalogues and matching with the KiDS imaging.
Appendix B: SOM quality control

The original quality control (Eq. 2) was calibrated for the spectroscopic calibration data of Hildebrandt et al. (2021) with $\sigma_{\text{mad}} = n\text{MAD} (\langle z_{\text{cal}} \rangle - \langle z_B \rangle) \approx 0.12$. Since this criterion may not be optimal for our new calibration dataset, we would in principle have needed to recalibrate the SOM cell rejection threshold to obtain the optimal trade-off between redshift bias and the size of the gold sample. It is, however, very challenging to simulate the photo-z errors and success rates of our new calibration dataset and therefore we chose to keep the threshold fixed for all gold sets.

Nevertheless, the addition of COSMOS2015 to the calibration data has a significant impact on the distribution of $\langle z_{\text{cal}} \rangle - \langle z_B \rangle$ for the fully trained cells of the SOM (Fig. B.1). While the distribution is almost symmetrical for the case of spec-z fiducial, it is skewed to positive values for spec-z+PAUS+COS15. The effect is more distinct for those cells that are occupied exclusively by the COSMOS2015 data than for those that exist in both gold samples (compare the orange dotted and black lines in Fig. B.1). This is a clear indication that much of the additional calibration data are faint objects for which the KiDS nine-band photo-z catastrophically underestimates the redshift obtained from the much deeper COSMOS2015 data. This effect is responsible for the reduced representation fraction in the first tomographic bin, as discussed in Sect. 5.1.1.

Fig. B.1. Comparison of the mean calibration sample redshift and KiDS nine-band photo-z in each SOM cell within the fiducial ±5$\sigma_{\text{mad}}$ interval. The coloured solid lines represent the fiducial (purple) and the full (orange) gold sample. The black line is the subset of SOM cells that belong to both gold samples, whereas the dashed orange line represents those cells that are occupied only by data from COSMOS2015.
Appendix C: Differences from the original KiDS-1000 analysis

The major difference between the analysis of the original KiDS-1000 data and this work is that our redshift calibration strategy focused exclusively on improving the SOM redshifts. We omitted revisiting the KiDS clustering redshifts, which Hildebrandt et al. (2021) used as a validation for the fiducial SOM redshift distributions. It is important to note, that the additional calibration data are only beneficial for the SOM calibration, since such inhomogeneous data, in particular photometric redshifts, cannot be applied easily in clustering redshifts.

Furthermore there are small differences in the cosmological analysis compared to A21, which are mostly related to small differences in the COSEBI data vector. We find that these differences originate from using a newer version of TreeCorr; however, it has no impact on our cosmological constraints which agree within the expected statistical variance of the sampler. Differences in the measured COSEBIs also propagate to the covariance matrix, resulting in sub-percent differences between our covariance and the one used by A21. Finally, we assumed 99% correlation between m-bias values per tomographic bin instead of 100% (perfect correlation) as in A21. The effect of this choice is negligible compared to those introduced by the variations in the COSEBI data vector.

We also tested whether the correction of the m-bias (Table 2) has a significant impact on the cosmological constraints in A21 by analysing the same data (‘Asgari new m-bias’) with otherwise identical parameters (except for the differences summarised above). Our results show that the recovered primary cosmological parameters, in particular $S_8 = 0.757^{+0.024}_{-0.023}$, are almost identical with those of our reanalysis of the KiDS-1000 data, performed with the old m-bias values (Table C.1). Given that the uncertainties of the m-biases are unchanged and the central value changes by less than $0.5\sigma$ in any of the tomographic bins, these results are as expected and have no further implication for the conclusions of A21.

Table C.1. Summary of the main cosmological parameter constraints (best fit and 68th-percentile PJ-HPD) from COSEBIs for Asgari et al. (2021) and our reanalysis with and without the corrected m-bias values.

<table>
<thead>
<tr>
<th>Sample</th>
<th>$\chi^2$</th>
<th>$A_{1A}$</th>
<th>$\Omega_m$</th>
<th>$\sigma_8$</th>
<th>$S_8$</th>
<th>$\Sigma_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asgari et al. (2021)</td>
<td>82.2</td>
<td>0.264$^{+0.424}_{-0.336}$</td>
<td>0.246$^{+0.0101}_{-0.0060}$</td>
<td>0.838$^{+0.140}_{-0.141}$</td>
<td>0.759$^{+0.024}_{-0.021}$</td>
<td>0.751$^{+0.024}_{-0.016}$</td>
</tr>
<tr>
<td>Asgari reanalysis</td>
<td>83.8</td>
<td>0.243$^{+0.370}_{-0.424}$</td>
<td>0.240$^{+0.101}_{-0.059}$</td>
<td>0.846$^{+0.146}_{-0.146}$</td>
<td>0.758$^{+0.029}_{-0.018}$</td>
<td>0.749$^{+0.025}_{-0.013}$</td>
</tr>
<tr>
<td>Asgari new m-bias</td>
<td>83.2</td>
<td>0.233$^{+0.406}_{-0.388}$</td>
<td>0.242$^{+0.104}_{-0.060}$</td>
<td>0.844$^{+0.114}_{-0.165}$</td>
<td>0.757$^{+0.022}_{-0.023}$</td>
<td>0.749$^{+0.024}_{-0.017}$</td>
</tr>
</tbody>
</table>

Notes. Additionally shown are the $\chi^2$ values (for 70.5 effective degrees of freedom) and $\Sigma_8 = \sigma_8(\Omega_m/0.3)^{\alpha}$ calculated for $\alpha_{1A} = 0.55$, determined from a fit to the posterior samples of the fiducial chain.
Appendix D: Marginal parameter constraints

Table D.1 lists an alternative representation of the parameter constraints of Table 3, showing the maximum of the marginal distribution and the associated 68th-percentile instead of the best fit and the 68th-percentile PJ-HPD.

Table D.1. Summary of the main cosmological parameter constraints (maximum of the marginal distribution and the associated 68th-percentile) from COSEBIIs for all gold samples and the comparison to Asgari et al. (2021) and Planck legacy (TT, TE, EE + lowE).

<table>
<thead>
<tr>
<th>Sample</th>
<th>$\Lambda$</th>
<th>$\Omega_m$</th>
<th>$\sigma_8$</th>
<th>$S_8$</th>
<th>$\Sigma_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>spec-z fiducial</td>
<td>0.394</td>
<td>0.278</td>
<td>0.734</td>
<td>0.745</td>
<td>0.740</td>
</tr>
<tr>
<td>spec-z+PAUS</td>
<td>0.108</td>
<td>0.238</td>
<td>0.800</td>
<td>0.745</td>
<td>0.736</td>
</tr>
<tr>
<td>spec-z+PAUS+COS15</td>
<td>0.440</td>
<td>0.206</td>
<td>0.843</td>
<td>0.750</td>
<td>0.739</td>
</tr>
<tr>
<td>spec-z nQ ≥ 4</td>
<td>0.088</td>
<td>0.203</td>
<td>0.877</td>
<td>0.758</td>
<td>0.746</td>
</tr>
<tr>
<td>only PAUS</td>
<td>0.078</td>
<td>0.210</td>
<td>0.879</td>
<td>0.753</td>
<td>0.740</td>
</tr>
<tr>
<td>only COS15</td>
<td>0.345</td>
<td>0.245</td>
<td>0.672</td>
<td>0.736</td>
<td>0.735</td>
</tr>
<tr>
<td>only PAUS+COS15</td>
<td>0.392</td>
<td>0.264</td>
<td>0.743</td>
<td>0.737</td>
<td>0.736</td>
</tr>
<tr>
<td>Asgari et al. (2021)</td>
<td>0.384</td>
<td>0.254</td>
<td>0.773</td>
<td>0.758</td>
<td>0.753</td>
</tr>
<tr>
<td>Planck legacy</td>
<td>0.136</td>
<td>0.210</td>
<td>0.813</td>
<td>0.836</td>
<td>0.838</td>
</tr>
</tbody>
</table>

Notes. Additionally shown is $\Sigma_8 = \sigma_8(\Omega_m/0.3)^{0.55}$ calculated for $\alpha_{sd} = 0.55$, determined from a fit to the posterior samples of the fiducial chain.
Appendix E: Goodness of fit

We find that the cosmological model fits to the new gold samples achieve a better goodness-of-fit ($\chi^2$ between 62 and 74, Table 3) than those to the original KiDS-1000 gold sample with $\chi^2 \approx 83$ as a reference. Since our methodology is identical in both cases, these differences stem from changes of the data vector that propagate also into the covariance matrix. We compared in particular the COSEBI data vectors with their respective best fit models for the spec-z fiducial and Asgari new $m$-bias (App. C) in Fig. E.1. Although a comparison by eye of the model and data is difficult due to the strong correlation between the different COSEBI modes, the most significant changes in the data vector occur between modes that involve bin one and two. In contrast to this the fitted models are very close in the first two bins and begin to deviate with increasing redshift.

These changes propagate to the $\chi^2$, which becomes apparent when summing the individual $\chi^2$-values in each panel of Fig. E.1 for the two left-most columns. While the sum over the remaining three columns (dominated by bin three, four, and five) is almost identical for spec-z fiducial and Asgari new $m$-bias, the first two columns fully account for the smaller goodness-of-fit of $\Delta \chi^2 \approx 20$. Since changing the $m$-bias seems to have a rather small impact on the $\chi^2$ (see Table C.1), this behaviour is most likely linked to the changes in the gold sample selection and redshift distributions, which are most significant for the first two bins. A21 report tensions between parameter constraints obtained from using only the second bin (including its cross-correlations) and the remainder of the data vector. A reduction of the $\chi^2$-value alone does not allow any conclusions on whether the improved calibration sample has an impact on the internal consistency of the cosmological analysis.

![Fig. E.1. Comparison of the COSEBI data vector and the best fit model for spec-z fiducial and the reanalysis of Asgari et al. (2021) with corrected $m$-bias. The triangular arrangement of all 15 combinations of tomographic bins are labelled with their individual $\chi^2$-values. The COSEBIs modes are significantly correlated, therefore their goodness-of-fit cannot be estimated by eye.](image-url)