A&A 647, A117 (2021)
https://doi.org/10.1051/0004-6361/202040237
© Euclid Collaboration 2021

Astronomy & Astrophysics

Euclid preparation

XI. Mean redshift determination from galaxy redshift probabilities for cosmic shear tomography

Euclid Collaboration: O. Ilbert[1], S. de la Torre[1], N. Martinet[1], A. H. Wright[2], S. Paltani[3], C. Laigle[4], I. Davidzon[5], E. Jullo[1], H. Hildebrandt[2], D. C. Masters[6], A. Amara[7], C. J. Conselice[8], S. Andreon[9], N. Auricchio[10], R. Azzollini[11], C. Baccigalupi[12,13,14,15], A. Balaguera-Antolínez[16,17], M. Baldi[10,18,19], A. Balestra[20], S. Bardelli[10], R. Bender[21,22], A. Biviano[12,15], C. Bodendorf[22], D. Bonino[23], S. Borgani[12,14,15,24], A. Boucaud[25], E. Bozzo[3], E. Branchini[26,27,28], M. Brescia[29], C. Burigana[30,31,32], R. Cabanac[33], S. Camera[23,34,35], V. Capobianco[23], A. Cappi[10,36], C. Carbone[37], J. Carretero[38], C. S. Carvalho[39], S. Casas[40], F. J. Castander[41,42], M. Castellano[28], G. Castignani[43], S. Cavuoti[29,44,45], A. Cimatti[18,46], R. Cledassou[47], C. Colodro-Conde[17], G. Congedo[48], L. Conversi[49,50], Y. Copin[51], L. Corcione[23], A. Costille[1], J. Coupon[3], H. M. Courtois[51], M. Cropper[11], J. Cuby[1], A. Da Silva[52,53], H. Degaudenzi[3], D. Di Ferdinando[19], F. Dubath[3], C. Duncan[54], X. Dupac[50], S. Dusini[55], A. Ealet[56], M. Fabricius[21,22], S. Farrens[40], P. G. Ferreira[54], F. Finelli[10,30], P. Fosalba[41,42], S. Fotopoulou[57], E. Franceschi[10], P. Franzetti[37], S. Galeotta[15], B. Garilli[37], W. Gillard[58], B. Gillis[48], C. Giocoli[10,18,19], G. Gozaliasl[59], J. Graciá-Carpio[22], F. Grupp[21,22], L. Guzzo[9,60,61], S. V. H. Haugan[62], W. Holmes[63], F. Hormuth[64], K. Jahnke[65], E. Keihanen[66], S. Kermiche[58], A. Kiessling[63], C. C. Kirkpatrick[66], M. Kunz[67], H. Kurki-Suonio[66], S. Ligori[23], P. B. Lilje[62], I. Lloro[68], D. Maino[37,60,61], E. Maiorano[10], O. Marggraf[69], K. Markovic[63], F. Marulli[10,18,19], R. Massey[70], M. Maturi[71,72], N. Mauri[18,19], S. Maurogordato[36], H. J. McCracken[4], E. Medinaceli[73], S. Mei[74,75], R. Benton Metcalf[18,73], M. Moresco[10,18], B. Morin[76,77], L. Moscardini[10,18,19], E. Munari[15], R. Nakajima[69], C. Neissner[38], S. Niemi[11], J. Nightingale[78], C. Padilla[38], F. Pasian[15], L. Patrizii[19], K. Pedersen[79], R. Pello[1], V. Pettorino[40], S. Pires[40], G. Polenta[80], M. Poncet[47], L. Popa[81], D. Potter[82], L. Pozzetti[10], F. Raison[22], A. Renzi[55,83], J. Rhodes[63], G. Riccio[29], E. Romelli[15], M. Roncarelli[10,18], E. Rossetti[18], R. Saglia[21,22], A. G. Sánchez[22], D. Sapone[84], P. Schneider[69], T. Schrabback[69], V. Scottez[4], A. Secroun[58], G. Seidel[65], S. Serrano[41,42], C. Sirignano[55,83], G. Sirri[19], L. Stanco[55], F. Sureau[40], P. Tallada Crespá[85], M. Tenti[19], H. I. Teplitz[6], I. Tereno[39,52], R. Toledo-Moreo[86], F. Torradeflot[85], A. Tramacere[3], E. A. Valentijn[87], L. Valenziano[10,19], J. Valiviita[66,88], T. Vassallo[21], Y. Wang[6], N. Welikala[48], J. Weller[21,22], L. Whittaker[8,89], A. Zacchei[15], G. Zamorani[10], J. Zoubian[58], and E. Zucca[10]

(Affiliations can be found after the references)

Received 24 December 2020 / Accepted 2 January 2021

ABSTRACT

The analysis of weak gravitational lensing in wide-field imaging surveys is considered to be a major cosmological probe of dark energy. Our capacity to constrain the dark energy equation of state relies on an accurate knowledge of the galaxy mean redshift $\langle z \rangle$. We investigate the possibility of measuring $\langle z \rangle$ with an accuracy better than $0.002\,(1+z)$ in ten tomographic bins spanning the redshift interval $0.2 < z < 2.2$, the requirements for the cosmic shear analysis of *Euclid*. We implement a sufficiently realistic simulation in order to understand the advantages and complementarity, as well as the shortcomings, of two standard approaches: the direct calibration of $\langle z \rangle$ with a dedicated spectroscopic sample and the combination of the photometric redshift probability distribution functions (zPDFs) of individual galaxies. We base our study on the Horizon-AGN hydrodynamical simulation, which we analyse with a standard galaxy spectral energy distribution template-fitting code. Such a procedure produces photometric redshifts with realistic biases, precisions, and failure rates. We find that the current *Euclid* design for direct calibration is sufficiently robust to reach the requirement on the mean redshift, provided that the purity level of the spectroscopic sample is maintained at an extremely high level of >99.8%. The zPDF approach can also be successful if the zPDF is de-biased using a spectroscopic training sample. This approach requires deep imaging data but is weakly sensitive to spectroscopic redshift failures in the training sample. We improve the de-biasing method and confirm our finding by applying it to real-world weak-lensing datasets (COSMOS and KiDS+VIKING-450).

Key words. dark energy – galaxies: distances and redshifts – methods: statistical

1. Introduction

Understanding the late, accelerated expansion of our Universe (Riess et al. 1998; Perlmutter et al. 1999) is one of the most important challenges in modern cosmology. Three leading hypotheses are: a modification of the laws of gravity, the introduction of a cosmological constant Λ in the equations describing the dynamics of our Universe, and the existence of a dark energy fluid with negative pressure. The last two hypotheses can be disentangled from each another by measuring the equation of state w of dark energy, which links its pressure to its density. Only the case $w = -1$ is compatible with a cosmological constant, and therefore any deviation from this value would invalidate the standard Λ cold dark matter (ΛCDM) model in favour of dark energy. This makes the precise measurement of w a key component of future cosmological experiments, such as *Euclid* (Laureijs et al. 2011), the *Vera C. Rubin* Observatory Legacy Survey of Space and Time (LSST; LSST Science Collaboration 2009), and the *Nancy Grace Roman* Space Telescope (Spergel et al. 2015).

Cosmic shear (see e.g., Kilbinger 2015; Mandelbaum 2018, for recent reviews), which is the coherent distortion of galaxy images by large-scale structures via weak gravitational lensing, offers the potential to measure w with great precision: The *Euclid* survey, in particular, aims at reaching 1% precision on the measurement of w using cosmic shear. One advantage of using lensing to measure w, compared to other probes, is that there exists a direct link between galaxy image geometrical distortions (i.e. the shear) and the gravitational potential of the intervening structures. When the shapes of, and distances to, galaxy sources are known, gravitational lensing allows one to probe the matter distribution of the Universe.

This discovery has led to the rapid growth of interest in using cosmic shear as a key cosmological probe, as evidenced by its successful application to several surveys. Constraints on the matter density parameter, $\Omega_{\rm m}$, and the normalisation of the linear matter power spectrum, σ_8, have been reported by the Canada-France-Hawaii Telescope Lensing Survey (CFHTLenS, Kilbinger et al. 2013), the Kilo Degree Survey (KiDS, Hildebrandt et al. 2017), the Dark Energy Survey (DES, Troxel et al. 2018), and the Hyper-Suprime Camera Survey (HSC, Hikage et al. 2019). These studies typically utilise so-called cosmic shear tomography (Hu 1999), whereby the cosmic shear signal is obtained by measuring the cross-correlation between galaxy shapes in different bins along the line of sight (i.e. tomographic bins). Large forthcoming surveys that also utilise cosmic shear tomography will enhance the precision of cosmological parameter measurements (e.g., $\Omega_{\rm m}$, σ_8, and w) while also enabling the measurement of any evolution in the dark energy equation of state, such as that parametrised by Caldwell et al. (1998): $w = w_0 + w_a\,(1 - a)$, where a is the scale factor.

Tomographic cosmic shear studies require accurate knowledge of the galaxy redshift distribution. The estimation and calibration of the redshift distribution has been identified as one of the most problematic tasks in current cosmic shear surveys since systematic bias in the distribution calibration directly influences the resulting cosmological parameter estimates. In particular, Joudaki et al. (2020) show that the $\Omega_{\rm m}$–σ_8 constraints from KiDS and DES can be fully reconciled under consistent redshift calibration, thereby suggesting that the different constraints from the two surveys can be traced back to differing methods of redshift calibration.

In tomographic cosmic shear, the signal is primarily sensitive to the average distance of sources within each bin. Therefore, for this purpose, the redshift distribution of an arbitrary galaxy sample can be characterised simply by its mean $\langle z\rangle$, defined as:

$$\langle z\rangle = \frac{\int_0^\infty z\,N(z)\,{\rm d}z}{\int_0^\infty N(z)\,{\rm d}z}, \tag{1}$$

where $N(z)$ is the true redshift distribution of the sample. Furthermore, in cosmic shear tomography, it is common to build the required tomographic bins using photo-z (see Salvato et al. 2019, for a review), which can be measured for large samples of galaxies with observations in only a few photometric bandpasses. However these photo-z are imperfect (due to, for example, photometric noise), resulting in tomographic bins whose true $N(z)$ extend beyond the bin limits. These 'tails' in the redshift distribution are important as they can significantly influence the distribution mean and provide sensitive information (Ma et al. 2006). For a *Euclid*-like cosmic shear survey, Laureijs et al. (2011) predict that the mean redshift $\langle z\rangle$ of each tomographic bin must be known with an accuracy better than $\sigma_{\langle z\rangle} = 0.002\,(1 + z)$ in order to meet the precision on w_0 ($\sigma_{w_0} = 0.015$) and w_a ($\sigma_{w_a} = 0.15$).

Given the importance of measuring the mean redshift for cosmic-shear surveys, numerous approaches have been devised in the last decade. A first family of methods, usually referred to as 'direct calibration', involves weighting a sample of galaxies with known redshifts such that they match the colour-magnitude properties of the target galaxy sample, thereby leveraging the relationship between galaxy colours, magnitudes, and redshifts to reconstruct the redshift distribution of the target sample (e.g., Lima et al. 2008; Cunha et al. 2009; Abdalla et al. 2008). A second approach is to utilise redshift probability distribution functions (zPDFs), obtained per target galaxy and subsequently stacked to reconstruct the target population $N(z)$. The galaxy zPDF is typically estimated by either model fitting or via machine learning. A third family of methods uses galaxy spatial information, specifically galaxy angular clustering, cross-correlating target galaxies with a large spec-z sample to retrieve the redshift distribution (e.g., Newman 2008; Ménard et al. 2013). New methods are continuously developed, for instance modelling galaxy populations and using forward modelling to match the data (Kacprzak et al. 2020).

In this paper, we evaluate our capacity to measure the mean redshift in each tomographic bin at the precision level required for *Euclid* based on realistic simulations. We base our study on a mock catalogue generated from the Horizon-AGN hydrodynamical simulation as described in Dubois et al. (2014) and Laigle et al. (2019). The advantage of this simulation is that the produced spectra encompass all the complexity of galaxy evolution, including rapidly varying star-formation histories, metallicity enrichment, mergers, and feedback from both supernovae and active galactic nuclei (AGN). By simulating galaxies with the imaging sensitivity expected for *Euclid*, we retrieve the photo-z with a standard template-fitting code, as done in existing surveys. Therefore, we produce photo-z with realistic biases, precisions, and failure rates, as shown in Laigle et al. (2019). The simulated galaxy zPDFs appear as complex as the ones observed in real data.

We further simulate realistic spectroscopic training samples with selection functions similar to those that are currently being acquired in preparation for *Euclid* and other dark energy experiments (Masters et al. 2017). We introduce possible incompleteness and failures to mimic those occurring in actual spectroscopic surveys.

We investigate two of the methods envisioned for the *Euclid* mission: direct calibration and zPDF combination. We also propose a new method to de-bias the zPDF based on Bordoloi et al. (2010). We quantify their performances in estimating the mean redshift of tomographic bins and isolate relevant factors that could impact our ability to fulfil the *Euclid* requirement. We also provide recommendations on the imaging depth and training sample necessary to achieve the required accuracy on $\langle z\rangle$.

Finally, we demonstrate the general utility of each of the methods presented here, not just to future surveys such as *Euclid* but also to current large imaging surveys. As an illustration, we apply these methods to the Cosmic Evolution Survey (COSMOS) survey and the fourth data release of KiDS (Kuijken et al. 2019).

The paper is organised as follows. In Sect. 2, we describe the mock *Euclid*-like catalogues generated from the Horizon-AGN hydrodynamical simulation. In Sect. 3, we test the precision reached on $\langle z\rangle$ when applying the direct calibration method. In Sect. 4, we measure the $\langle z\rangle$ in each tomographic bin using the zPDF de-biasing technique. We discuss the advantages and limitations of both methods in Sect. 5. We apply these methods to the KiDS and COSMOS dataset in Sect. 6. Finally, we summarise our findings and provide closing remarks in Sect. 7.

2. A mock *Euclid* catalogue

In this section, we present the mock *Euclid* catalogue used in this analysis, which is constructed from the Horizon-AGN hydrodynamical simulated lightcone and includes photometry and photometric redshift information. A full description of this mock catalogue can be found in Laigle et al. (2019). Here we summarise its main features and discuss the construction of several simulated spectroscopic samples, which reproduce a number of expected spectroscopic selection effects.

2.1. Horizon-AGN simulation

Horizon-AGN is a cosmological hydrodynamical simulation that was run in a simulation box of $100\,h^{-1}$ Mpc per side and with a dark matter mass resolution of $8\times10^7\,M_\odot$ (Dubois et al. 2014). A flat ΛCDM cosmology with $H_0 = 70.4\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, $\Omega_\mathrm{m} = 0.272$, $\Omega_\Lambda = 0.728$, and $n_\mathrm{s} = 0.967$ (compatible with WMAP-7, Komatsu et al. 2011) is assumed. Gas evolution is followed on an adaptive mesh, whereby an initial coarse 1024^3 grid is refined down to 1 physical kiloparsec. The refinement procedure leads to a typical number of 6.5×10^9 gas resolution elements (called leaf cells) in the simulation at $z = 1$. Following Haardt & Madau (1996), heating of the gas by a uniform ultraviolet background radiation field takes place after $z = 10$. Gas in the simulation is able to cool down to temperatures of 10^4 K through H and He collision and with a contribution from metals as tabulated in Sutherland & Dopita (1993). Gas is converted into stellar particles in regions where the gas particle number density surpasses $n_0 = 0.1\,\mathrm{H\,cm^{-3}}$, following a Schmidt law, as explained in Dubois et al. (2014). Feedback from stellar winds and supernovae (both types Ia and II) are included in the simulation, and it comprises mass, energy, and metal releases. Black holes (BHs) in the simulation can grow by gas accretion, at a Bondi accretion rate that is capped at the Eddington limit, and are able to coalesce when they form a sufficiently tight binary. They release energy in either the quasar or radio (i.e. heating or jet) mode, when the accretion rate is respectively above or below one percent of the Eddington ratio. The efficiency of these energy release modes is tuned to match the observed BH-galaxy scaling relation at $z = 0$ (see Dubois et al. 2012, for more details).

The simulation lightcone was extracted as described in Pichon et al. (2010). Particles and gas leaf cells were extracted at each time step depending on their proper distance to the observer at the origin. In total, the lightcone contains roughly 22 000 portions of concentric shells, which are taken from about 19 replications of the Horizon-AGN box up to $z = 4$. We restricted ourselves to the central 1 deg^2 of the lightcone. Laigle et al. (2019) extracted a galaxy catalogue from the stellar particle distribution using the AdaptaHOP halo finder (Aubert et al. 2004), where galaxy identification is based exclusively on the local stellar particle density. Only galaxies with stellar masses $M_\star > 10^9\,M_\odot$ (which corresponds to around 500 stellar particles) are kept in the final catalogue, resulting in more than 7×10^5 galaxies in the redshift range $0 < z < 4$, with a spatial resolution of 1 kpc.

A full description of the per-galaxy spectral energy distribution (SED) computation within Horizon-AGN is presented in Laigle et al. (2019)[1]; in the following, we only summarise the key details of the SED construction process. Each stellar particle in the simulation is assumed to behave as a single stellar population, and its contribution to the galaxy spectrum is generated using the stellar population synthesis models from Bruzual & Charlot (2003), assuming a Chabrier (2003) initial mass function. As each galaxy is composed of a large number of stellar particles, the galaxy SEDs therefore naturally capture the complexities of unique star-formation and chemical enrichment histories. Additionally, dust attenuation is also modelled for each star particle individually, using the mass distribution of the gas-phase metals as a proxy for the dust distribution and adopting a constant dust-to-metal mass ratio. Dust attenuation (neglecting scattering) is therefore inherently geometry-dependent in the simulation. Finally, the absorption of SED photons by the intergalactic medium (i.e. H I absorption in the Lyman series) is modelled along the line of sight to each galaxy using our knowledge of the gas density distribution in the lightcone. This, therefore, introduces variation into the observed intergalactic absorption across individual lines of sight. Flux contamination by nebular emission lines is not included in the simulated SEDs. While emission lines could add some complexity to a galaxy's photometry, their contribution can be modelled in a template-fitting code. Moreover, their impact is mostly crucial at high redshifts (Schaerer & de Barros 2009) and when using medium bands (e.g., Ilbert et al. 2009).

Kaviraj et al. (2017) compare the global properties of the simulated galaxies with statistical measurements available in the literature (as the luminosity functions, the star-forming main sequence, or the mass functions). They find an overall fairly good agreement with observations. Still, the simulation overpredicts the density of low-mass galaxies, and the median specific star-formation rate falls slightly below the literature results, a common trend in current simulations.

2.2. Simulation of Euclid photometry and photometric redshifts

As described in Laureijs et al. (2011), the *Euclid* mission will measure the shapes of about 1.5 billion galaxies over 15 000 deg^2. The visible (VIS) instrument will obtain images taken in one very broad filter (VIS), spanning 3500 Å. This filter allows extremely efficient light collection and will enable the VIS instrument to measure the shapes of galaxies as faint as 24.5 mag with high precision. The near-infrared spectrometer and photometer (NISP) instrument will produce images in three near-infrared (NIR) filters. In addition to these data, *Euclid* satellite observations are expected to be complemented by large samples of ground-based imaging, primarily in the optical, to assist the measurement of photo-z.

Euclid imaging has an expected sensitivity, over 15 000 deg^2, of 24.5 mag (at 10σ) in the VIS band, and 24 mag (at 5σ) in each of the Y, J, and H bands (Laureijs et al. 2011). We associate the *Euclid* imaging with two possible ground-based visible imaging datasets, which correspond to two limiting cases for photo-z estimation performance. The first is DES/*Euclid*. As a demonstration of photo-z performance when combining *Euclid* with a

[1] Horizon-AGN photometric catalogues and SEDs can be downloaded from https://www.horizon-simulation.org/data.html

considerably shallower photometric dataset, we combined our *Euclid* photometry with that from the DES (Abbott et al. 2018). The DES imaging is taken in the g, r, i, and z filters, at 10σ sensitivities of 24.33, 24.08, 23.44, and 22.69, respectively.

The second is LSST/*Euclid*. As a demonstration of photo-z performance when combining *Euclid* with a considerably deeper photometric dataset, we combined our *Euclid* photometry with that from the *Vera C. Rubin* Observatory LSST (LSST Science Collaboration 2009). The LSST imaging will be taken in the u, g, r, i, z, and y filters, at 5σ (point source, full depth) sensitivities of 26.3, 27.5, 27.7, 27.0, 26.2, and 24.9, respectively.

The DES imaging is completed and meets these expected sensitivities. Conversely, LSST will not reach the quoted full depth sensitivities before its tenth year of operation (i.e. starting in 2021), and even then it is possible that the northern extension of LSST might not reach the same depth. Still, LSST will already be extremely deep after two years of operation, being only 0.9 mag shallower than the final expected sensitivity (Graham et al. 2020). Therefore, these two cases (and their assumed sensitivities) should comfortably encompass the possible photo-z performance of any future combined optical and *Euclid* photometric dataset.

In order to generate the mock photometry in each of the *Euclid*, DES, and LSST surveys, each galaxy SED is first 'observed' through the relevant filter response curves. In each photometric band, we generated Gaussian distributions of the expected signal-to-noise ratios (S/Ns) as a function of magnitude, given both the depth of the survey and the typical S/N-magnitude relation (in the same wavelength range) (see Appendix A in Laigle et al. 2019). We then used these distributions, per filter, to assign each galaxy a S/N (based on its magnitude). The S/N of each galaxy determines its 'true' flux uncertainty, which is then used to perturb the photometry (assuming Gaussian random noise) and produce the final flux estimate per source. This process was then repeated for all desired filters.

The galaxy photo-z were derived in the same manner as with real-world photometry. We used the method detailed in Ilbert et al. (2013), which is based on the template-fitting code `LePhare` (Arnouts et al. 2002; Ilbert et al. 2006). We adopted a set of 33 templates from Polletta et al. (2007), which was complemented with templates from Bruzual & Charlot (2003). Two dust attenuation curves were considered (Prevot et al. 1984; Calzetti et al. 2000), allowing for a possible bump at 2175 Å. Neither emission lines nor the adaptation of the zero-points were considered since they were not included in the simulated galaxy catalogue. The full redshift likelihood, $\mathcal{L}(z)$, is stored for each galaxy, and the photo-z point-estimate, $z_{\rm p}$, is defined as the median of $\mathcal{L}(z)$[2]. The distributions of (derived) photometric redshift versus (intrinsic) spectroscopic redshift for mock galaxies (in both our DES/*Euclid* and LSST/*Euclid* configurations) are shown in Fig. 1. Several examples of redshift likelihoods are shown in Fig. 2. We can see realistic cases with multiple modes in the distribution, as well as asymmetric distributions around the main mode. The photo-z used to select galaxies within the tomographic bins are indicated by the magenta lines, which can differ significantly from the spec-z (green lines).

We wished to remove galaxies with a broad likelihood distribution (i.e. galaxies with truly uncertain photo-z) from our sample. In practice, we approximated the breadth of the likelihood distribution using the photo-z uncertainties produced by the template-fitting procedure to clean the sample. `LePhare` produces a redshift confidence interval $[z_{\rm p}^{\rm min}, z_{\rm p}^{\rm max}]$, per source, which encompasses 68% of the redshift probability around $z_{\rm p}$. We removed galaxies with $\max(z_{\rm p} - z_{\rm p}^{\rm min}, z_{\rm p}^{\rm max} - z_{\rm p}) > 0.3$, which we denote $\sigma_{z_{\rm p}} > 0.3$ in the following for simplicity. We investigate the impact of this choice on the number of galaxies available for cosmic shear analyses and quantify the impact of relaxing this limit in Sect. 5.2.

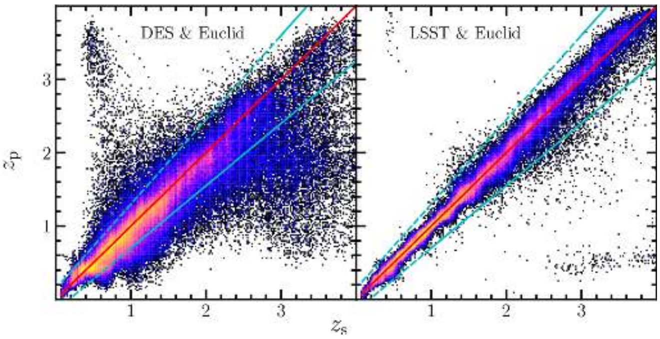

Fig. 1. Comparison between the photometric redshifts ($z_{\rm p}$) and spectroscopic redshifts ($z_{\rm s}$) for the simulated Horizon-AGN galaxy sample. Each panel shows a two-dimensional histogram with logarithmic colour scaling and is annotated with both the 1:1 equivalence line (red) and the $|z_{\rm p} - z_{\rm s}| = 0.15\,(1 + z_{\rm s})$ outlier thresholds (blue) for reference. Photometric redshifts are computed using both DES/*Euclid* (*left*) and LSST/*Euclid* (*right*) simulated photometry, assuming a *Euclid*-based magnitude-limited sample with VIS < 24.5.

Finally, we generated 18 photometric noise realisations of the mock galaxy catalogue. While the intrinsic physical properties of the simulated galaxies remain the same under each of these realisations, the differing photometric noise allows us to quantify the role of photometric noise alone on our estimated $\langle z \rangle$. We only adopted 18 realisations due to computational limitations; however, our results are stable to the addition of more realisations.

2.3. Definition of the target photometric sample and the spectroscopic training samples

All redshift calibration approaches discussed in this paper utilise a spec-z training sample to estimate the mean redshift of a target photometric sample. In practice, such a spectroscopic training sample is rarely a representative subset of the target photometric sample; rather, it is often composed of bluer and brighter galaxies. Therefore, to properly assess the performance of our tested approaches, we had to ensure that the simulated training sample is distinct from the photometric sample. To do this, we separated the Horizon-AGN catalogue into two equally sized subsets: We defined the first half of the photometric catalogue as our as target sample and drew variously defined spectroscopic training samples from the second half of the catalogue. We tested each of our calibration approaches with three spectroscopic training samples designed to mimic different spectroscopic selection functions: (1) a uniform training sample; (2) a self-organising map-based training sample; and (3) a COSMOS-like training sample.

The uniform training sample is the simplest, most idealised training sample possible. We sampled 1000 galaxies with VIS < 24.5 mag (i.e. the same magnitude limit as in the target sample) in each tomographic bin, independently of all other properties. While this sample is ideal in terms of representation, the sample

[2] The median of $\mathcal{L}(z)$ could differ from the peak of $\mathcal{L}(z)$ or from the redshift corresponding to the minimum χ^2, especially for ill-defined likelihoods.

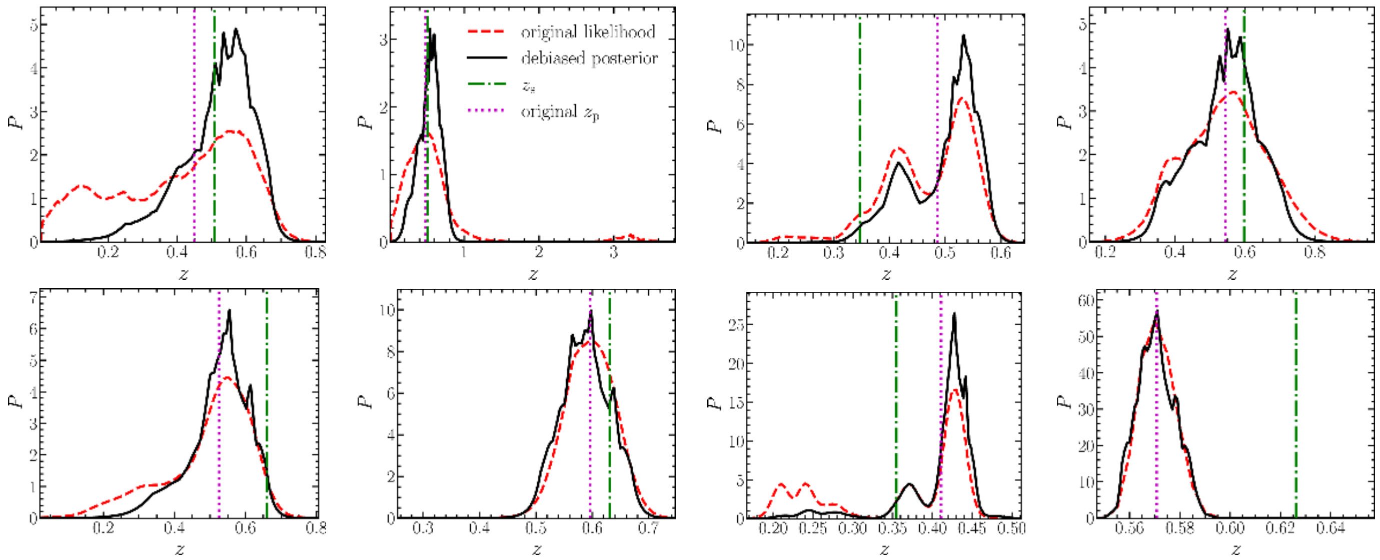

Fig. 2. Examples of galaxy likelihood $\mathcal{L}(z)$ (dashed red lines) and de-biased posterior distributions (solid black lines). The spec-z (photo-z) are indicated with dotted green (magenta) lines. These galaxies are selected in the tomographic bin $0.4 < z_{\rm p} < 0.6$ for the DES/*Euclid* (*top panels*) and LSST/*Euclid* (*bottom panels*) configurations. These likelihoods are not a random selection of sources, but illustrate the variety of likelihoods present in the simulations.

size was set to mimic a realistic training sample that could be obtained from dedicated ground-based spectroscopic follow-up of a *Euclid*-like target sample.

Our second training sample follows the current *Euclid* baseline to build a training sample. Masters et al. (2017) have endeavoured to construct a spectroscopic survey, the Complete Calibration of the Colour-Redshift Relation survey (C3R2), which completely samples the colour and magnitude space of cosmic shear target samples. This sample is currently being assembled by combining data from ESO and Keck facilities (Masters et al. 2019; Guglielmo et al. 2020). The target selection is based on an unsupervised machine-learning technique, the self-organising map (SOM, Kohonen 1982), which they use to define a spectroscopic target sample that is representative in terms of the galaxy colours of the *Euclid* cosmic shear sample. The SOM allows a projection of a multi-dimensional distribution onto a lower two-dimensional map. The utility of the SOM lies in its preservation of higher-dimensional topology: Neighbouring objects in the multi-dimensional space fall within similar regions of the resulting map. This allows the SOM to be utilised as a multi-dimensional clustering tool, whereby discrete map cells associate sources within discrete voxels in the higher-dimensional space. We used the method from Davidzon et al. (2019) to construct a SOM, which involves projecting observed (i.e. noisy) colours of the mock catalogue onto a map of 6400 cells (with dimension 80 × 80). We constructed our SOM using the LSST/*Euclid* simulated colours, implicitly assuming that the spec-z training sample is defined using deep calibration fields. If the flux uncertainty is too large ($\Delta m_i^x > 0.5$, for object i in filter x), the observed magnitude is replaced by that predicted from the best-fit SED template, which is estimated while preparing the SOM input catalogue. This procedure allows us to retain sources that have non-detections in some photometric bands. We then constructed our SOM-based training sample by randomly selecting $N_{\rm train}$ galaxies from each cell in the SOM. The C3R2 expects to have ≥1 spectroscopic galaxies per SOM cell available for calibration by the time the *Euclid* mission is active. For our default SOM coverage, we invoked a slightly more idealised situation of two galaxies per cell and we imposed that these two galaxies belong to the considered tomographic bin. This procedure ensures that all cells are represented in the spectroscopy. In reality, a fraction of cells will likely not contain spectroscopy. However, when treated correctly, such misrepresented cells act only to decrease the target sample number density and do not bias the resulting redshift distribution mean estimates (Wright et al. 2020). We therefore expect that this idealised treatment will not produce results that are overly optimistic.

Finally, the COSMOS-like training sample mimics a typical heterogeneous spectroscopic sample, which is currently available in the COSMOS field. We first simulated the zCOSMOS-like spectroscopic sample (Lilly et al. 2007), which consists of two distinct components: a bright and a faint survey. The zCOSMOS-Bright sample was selected such that it contains only galaxies at $z < 1.2$, while the zCOSMOS-Faint sample contains only galaxies at $z > 1.7$ (with a strong bias towards selecting star-forming galaxies). To mimic these selections, we constructed a mock sample whereby half of the sources are brighter than $i = 22.5$ (the bright sample) and half of the galaxies reside at $1.7 < z < 2.4$ with $g < 25$ (the faint sample). We then added to this compilation a sample of 2000 galaxies that were randomly selected at $i < 25$, mimicking the low-z VIMOS Ultra Deep Survey (VUDS) sample (Le Fèvre et al. 2015), as well as a sample of 1000 galaxies randomly selected at $0.8 < z < 1.6$ with $i < 24$, mimicking the sample from Comparat et al. (2015). By construction, this final spectroscopic redshift compilation exhibits low representation of the photometric target sample in the redshift range $1.3 < z < 1.7$.

Overall, our three training samples exhibit (by design) differing redshift distributions and galaxy number densities. We investigate the sensitivity of the estimated $\langle z \rangle$ on the size of the training sample in Sect. 5.3.

3. Direct calibration

Direct calibration is a fairly straightforward method that can be used to estimate the mean redshift of a photometric galaxy sample, and it is currently the baseline method planned for *Euclid* cosmic shear analyses. In this section, we describe our implementation of the direct calibration method, apply this method to our various spectroscopic training samples, and report the resulting accuracy of our redshift distribution mean estimates.

3.1. Implementation for the different training samples

Given our different classes of training samples, we were able to implement slightly different methods of direct calibration. We detail here how the implementation of direct calibration differs for each of our three spectroscopic training samples.

The uniform sample. In the case where the training sample is known to uniformly sparse-sample the target galaxy distribution, an estimate of $\langle z \rangle$ can be approximated by simply computing the mean redshift of the training sample.

The SOM sample. By construction, the SOM training sample uniformly covers the full n-dimensional colour space of the target sample. The method relies on the assumption that galaxies within a cell share the same redshift (Masters et al. 2015), which can be labelled with the training sample. Therefore, we can estimate the mean redshift of the target distribution $\langle z \rangle$ by simply calculating the weighted mean of each cell's average redshift, where the weight is the number of target galaxies per cell,

$$\langle z \rangle = \frac{1}{N_{\rm t}} \sum_{i=1}^{N_{\rm cells}} \left\langle z^i_{\rm train} \right\rangle N_i, \tag{2}$$

where the sum runs over the $i \in [1, N_{\rm cells}]$ cells in the SOM, $\left\langle z^i_{\rm train} \right\rangle$ is the mean redshift of the training spectroscopic sources in cell i, N_i is the number of target galaxies (per tomographic bin) in cell i, and $N_{\rm t}$ is the total number of target galaxies in the tomographic bin. A shear weight associated with each galaxy can be introduced in this equation (e.g., Wright et al. 2020). As described in Sect. 2.3, our SOM was consistently constructed by training on LSST/*Euclid* photometry, even when studying the shallower DES/*Euclid* configuration. We adopted this strategy since the training spectroscopic samples in *Euclid* will be acquired in calibration fields (e.g., Masters et al. 2019) with deep dedicated imaging. This assumption implies that the target distribution $\langle z \rangle$ is estimated exclusively in these calibration fields, which are covered with photometry from both our shallow and deep setups, and therefore increases the influence of sample variance on the calibration.

The COSMOS-like sample. Applying direct calibration to a heterogeneous training sample is less straightforward than in the above cases as the training sample is not representative of the target sample in any respect. Weighting of the spectroscopic sample, therefore, must correct for the mix of spectroscopic selection effects present in the training sample, as a function of magnitude (from the various magnitude limits of the individual spectroscopic surveys), colour (from their various preselections in colour and spectral type), and redshift (from dedicated redshift preselection, such as that in zCOSMOS-Faint). Such a weighting scheme can be established efficiently with machine-learning techniques such as the SOM. To perform this weighting, we trained a new SOM using all the information that has the potential to correct for the selection effects present in our heterogeneous training sample: apparent magnitudes, colours, and template-based photo-z. We created this SOM using only the galaxies from the COSMOS-like sample that belong to the considered tomographic bin and reduced the size of the map to 400 cells (20 × 20, because the tomographic bin itself spans a smaller colour space). Finally, we projected the target sample into the SOM and derived weights for each training sample galaxy, such that they reproduce the per-cell density of target sample galaxies. This process follows the same weighting procedure as Wright et al. (2020), who extended the direct calibration method of Lima et al. (2008) to include source groupings defined via the SOM. In this method, the estimate of $\langle z \rangle$ is also inferred using Eq. (2).

3.2. Results

We applied the direct calibration technique to the mock catalogue, which was split into ten tomographic bins spanning the redshift interval $0.2 < z_{\rm p} < 2.2$. To construct the samples within each tomographic bin, the training and target samples are selected based on their best-estimate photo-z, $z_{\rm p}$. We quantified the performance of the redshift calibration procedure using the measured bias in $\langle z \rangle$, defined as

$$\Delta_{\langle z \rangle} = \frac{\langle z \rangle - \langle z \rangle^{\rm true}}{1 + \langle z \rangle^{\rm true}} \tag{3}$$

and evaluated over the target sample. We present the values of $\Delta_{\langle z \rangle}$ that we obtained with direct calibration for each of the ten tomographic bins in Fig. 3. The figure shows, per tomographic bin, the population mean (points) and 68% population scatter (errorbars) of $\Delta_{\langle z \rangle}$ over the 18 photometric noise realisations of our simulation. The solid lines and yellow region indicate the $|\Delta_{\langle z \rangle}| \leq 2 \times 10^{-3}$ requirement stipulated by the *Euclid* mission. Given our limited number of photometric noise realisations, estimating the population mean and scatter directly from the 18 samples is not sufficiently robust for our purposes. We thus used maximum likelihood estimation, assuming Gaussianity of the $\Delta_{\langle z \rangle}$ distribution, to determine the underlying population mean and the scatter. We define these underlying population statistics as $\mu_{\Delta z}$ and $\sigma_{\Delta z}$ for the mean and the scatter, respectively.

We find that, when using a uniform or SOM training sample, direct calibration is consistently able to recover the target sample mean redshift to $|\mu_{\Delta z}| < 2 \times 10^{-3}$. In the case of the shallow DES/*Euclid* configuration, however, the scatter $\sigma_{\Delta z}$ exceeds the *Euclid* accuracy requirement in the highest and lowest tomographic bins. The DES/*Euclid* configuration is, therefore, technically unable to meet the *Euclid* precision requirement on $\langle z \rangle$ in the extreme bins. In the LSST/*Euclid* configuration, conversely, the precision and accuracy requirements are both consistently satisfied. We hypothesise that this difference stems from the deeper photometry having higher discriminatory power in the tomographic binning itself: The $N(z)$ distribution for each tomographic bin is intrinsically broader for bins defined with shallow photometry and therefore has the potential to demonstrate greater complexity (such as colour-redshift degeneracies), which reduces the effectiveness of direct calibration.

The direct calibration with the SOM relies on the assumption that galaxies within a cell share the same redshift (Masters et al. 2015). Noise and degeneracies in the colour-redshift space introduce a redshift dispersion within the cell that impacts the accuracy of $\langle z \rangle$. Even with the diversity of SEDs generated with Horizon-AGN, and introducing noise into the photometry, we find that the direct calibration with a SOM sample is sufficient to reach the *Euclid* requirement.

We find that the COSMOS-like training sample is unable to reach the required accuracy of *Euclid*. This behaviour is somewhat expected since the COSMOS-like sample contains selection effects that are not cleanly accessible to the direct calibration weighting procedure. The mean redshift is particularly biased in

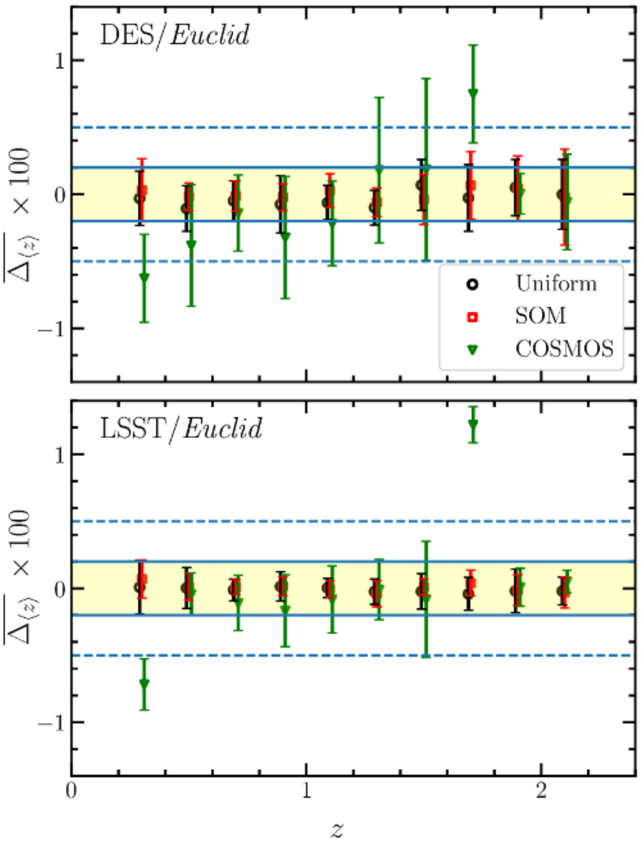

Fig. 3. Bias on the mean redshift (see Eq. (3)) averaged over the 18 photometric noise realisations. The mean redshifts are measured using the direct calibration approach. The tomographic bins are defined using the DES/*Euclid* and LSST/*Euclid* photo-z in the *top* and *bottom panels*, respectively. The yellow region represents the *Euclid* requirement at $0.002\,(1 + z)$ for the mean redshift accuracy, and the dashed blue lines correspond to a bias of $0.005\,(1 + z)$. The symbols represent the results obtained with different training samples: (a) uniformly selecting 1000 galaxies per tomographic bin (black circles); (b) selecting two galaxies per cell in the SOM (red squares); and (c) selecting a sample that mimics real spectroscopic survey compilations in the COSMOS field (green triangles).

the bin $1.6 < z < 1.8$, where there is a dearth of spectra; the Comparat et al. (2015) sample is limited to $z < 1.6$, while the zCOSMOS-Faint sample resides exclusively at $z > 1.7$, thereby leaving the range $1.6 < z < 1.7$ almost entirely unrepresented. In this circumstance, our SOM-based weighting procedure is insufficient to correct for the heterogeneous selection, leading to bias. This is typical in cases where the training sample is missing certain galaxy populations that are present in the target sample (Hartley et al. 2020). We note, though, that it may be possible to remove some of this bias via careful quality control during the direct calibration process, as demonstrated in Wright et al. (2020). Whether such quality control would be sufficient to meet the *Euclid* requirements, however, is uncertain.

We note that, although we are utilising photometric noise realisations in our estimates of $\langle z \rangle$, the underlying mock catalogue remains the same. As a result, our estimates of $\mu_{\Delta z}$ and $\sigma_{\Delta z}$ are not impacted by sample variance. In reality, sample variance affects the performance of the direct calibration, particularly when assuming that the training sample is directly representative of the target distribution (as we do with our uniform training sample). For fields smaller than $2\,\mathrm{deg}^2$, Bordoloi et al. (2010) showed that Poisson noise dominates over sample variance (in mean redshift estimation) when the training sample consists of fewer than 100 galaxies. Above this size, sample variance dominates the calibration uncertainty. This means that, in order to generate an unbiased estimate of $\langle z \rangle$ using a uniform sample of 1000 galaxies, a minimum of ten fields of $2\,\mathrm{deg}^2$ would need to be surveyed.

The SOM approach is less sensitive to sample variance, as over-densities (and under-densities) in the target sample population relative to the training sample are essentially removed in the weighting procedure (provided that the population is present in the training sample, Lima et al. 2008; Wright et al. 2020). In the cells corresponding to this over-represented target population, the relative importance of training sample redshifts will be similarly up-weighted, thereby removing any bias in the reconstructed $N(z)$. Therefore, sample variance should only have a weak impact on the global derived $N(z)$ in this method. Nonetheless, sample variance may still be problematic if, for example, under-densities result in entire populations being absent from the training sample.

Finally, it is worth emphasising that these results are obtained assuming a perfect knowledge of training set redshifts. We study the impact of failures in spectroscopic redshift estimation in Sect. 5.

4. Estimator based on redshift probabilities

In this section, we present another approach to redshift distribution calibration that uses the information contained in the galaxy zPDF, which is available for each individual galaxy of the target sample. Photometric redshift estimation codes typically provide approximations to this distribution based solely on the available photometry of each source. We study the performance of methods utilising this information in the context of *Euclid* and test a method to de-bias the zPDF.

4.1. Formalism

Given the relationship between galaxy magnitudes and colours (denoted $\boldsymbol{o}$) and redshift z, one can utilise the conditional probability $p(z|\boldsymbol{o})$ to estimate the true redshift distribution $N(z)$ using an estimator such as that from Sheth (2007), Sheth & Rossi (2010):

$$N(z) = \int N(\boldsymbol{o})\, p(z|\boldsymbol{o})\, \mathrm{d}\boldsymbol{o} = \sum_{i}^{N_\mathrm{t}} p_i(z|\boldsymbol{o}), \tag{4}$$

where $N(\boldsymbol{o})$ is the joint n-dimensional distribution of colours and magnitudes. As made explicit in the above equation, the $N(z)$ estimator simply reduces to the sum of the individual (per-galaxy) conditional redshift probability distributions, $p_i(z|\boldsymbol{o})$. A shear weight associated with each galaxy can be introduced in this equation (e.g., Wright et al. 2020). It is worth noting that this summation over conditional probabilities is ideologically similar to the summation of SOM-cell redshift distributions presented previously; in both cases, one effectively builds an estimate of the probability $p(z|\boldsymbol{o})$ and uses this to estimate $\langle z \rangle$. Indeed, it is clear that the SOM-based estimate of $\langle z \rangle$ presented in Eq. (2) does in fact follow directly from Eq. (4).

Generally, photometric redshift codes output a normalised likelihood function that provides the probability of the observed photometry if given the true redshift, $\mathcal{L}(\boldsymbol{o}|z)$, or sometimes the posterior probability distribution, $\mathcal{P}(z|\boldsymbol{o})$ (e.g., Benítez 2000;

Bolzonella et al. 2000; Arnouts et al. 2002; Cunha et al. 2009). These two probability distribution functions are related through the Bayes' theorem as

$$\mathcal{P}(z|\boldsymbol{o}) \propto \mathcal{L}(\boldsymbol{o}|z)\,\mathrm{Pr}(z), \tag{5}$$

where $\mathrm{Pr}(z)$ is the prior probability.

Photometric redshift methods that invoke template fitting, such as the `LePhare` photo-z estimation code, generally explore the likelihood of the observed photometry given a range of theoretical templates, T, and true redshifts, $\mathcal{L}(\boldsymbol{o}|T, z)$. The full likelihood, $\mathcal{L}(\boldsymbol{o}|z)$, is then obtained by marginalising over the template set:

$$\mathcal{L}(\boldsymbol{o}|z) = \sum_T \mathcal{L}(\boldsymbol{o}|T, z). \tag{6}$$

In the full Bayesian framework, however, we are instead interested in the posterior probability, rather than the likelihood. In the formulation of this posterior, we first made explicit the dependence between galaxy colours, $\boldsymbol{c}$, and magnitude in one (reference) band, m_0: $\boldsymbol{o} = \{\boldsymbol{c}, m_0\}$. Following Benítez (2000), we were then able to define the posterior probability distribution function,

$$\mathcal{P}(z|\boldsymbol{c}, m_0) \propto \sum_T \mathcal{L}(\boldsymbol{c}|T, z)\,\mathrm{Pr}(z|T, m_0)\,\mathrm{Pr}(T|m_0), \tag{7}$$

where $\mathrm{Pr}(z|T, m_0)$ is the prior conditional probability of redshift given a particular galaxy template and reference magnitude and $\mathrm{Pr}(T|m_0)$ is the prior conditional probability of each template at a given reference magnitude. Under the approximation that the redshift distribution does not depend on the template, and that the template distribution is independent of the magnitude (i.e. the luminosity function does not depend on the SED type), one obtains

$$\mathcal{P}(z|\boldsymbol{c}, m_0) \propto \sum_T \mathcal{L}(\boldsymbol{c}|T, z)\,\mathrm{Pr}(z|m_0), \tag{8}$$

$$\propto \mathcal{L}(\boldsymbol{c}|z)\,\mathrm{Pr}(z|m_0). \tag{9}$$

Adding the template dependency in the prior would improve our results, but this is impractical with the iterative method presented in Sect. 4 given the size of our sample.

The posterior probability $\mathcal{P}(z|\boldsymbol{o})$ is a photometric estimate of the true conditional redshift probability $p(z|\boldsymbol{o})$ in Eq. (4), and thus we are able to estimate the target sample $N(z)$ via the stacking of the individual galaxy posterior probability distributions,

$$N(z) = \sum_i^{N_\mathrm{t}} \mathcal{P}_i(z|\boldsymbol{o}), \tag{10}$$

and therefore

$$\langle z\rangle = \frac{\int z\left[\sum_i^{N_\mathrm{t}} \mathcal{P}_i(z|\boldsymbol{o})\right]\mathrm{d}z}{\int\left[\sum_i^{N_\mathrm{t}} \mathcal{P}_i(z|\boldsymbol{o})\right]\mathrm{d}z}. \tag{11}$$

4.2. Initial results

In this analysis, we used the `LePhare` code, which outputs $\mathcal{L}(\boldsymbol{o}|z)$ for each galaxy as defined in Eq. (6). The redshift distribution (and thereafter its mean) are obtained by summing galaxy posterior probabilities, which are derived as in Eq. (9). This raises, however, an immediate concern: In order to estimate the $N(z)$ using the per-galaxy likelihoods, we require a prior distribution of magnitude-dependant redshift probabilities, $\mathrm{Pr}(z|m_0)$, which naturally requires knowledge of the magnitude-dependent redshift distribution.

We tested the sensitivity of our method to this prior choice by considering priors of two types: a (formally improper) 'flat prior' with $\mathrm{Pr}(z|m_0) = 1$; and a 'photo-z prior' that is constructed by normalising the redshift distribution, estimated per magnitude bin, as obtained by summation over the likelihoods (following Brodwin et al. 2006). Formally, this photo-z prior is defined as

$$\mathrm{Pr}(z|m_0) = \sum_i^{N_\mathrm{t}} \mathcal{L}_i(\boldsymbol{o}|z)\,\Theta(m_{0,i}|m_0), \tag{12}$$

where $\Theta(m_{0,i}|m_0)$ is unity if $m_{0,i}$ is inside the magnitude bin centred on m_0 and zero otherwise, and N_t is the number of galaxies in the tomographic bin.

We estimated $\langle z\rangle$ in the previously defined tomographic bins using Eq. (11). In the upper-left panel of Fig. 4, we show estimated (and true) $N(z)$ for one tomographic bin with $1.2 < z_\mathrm{p} < 1.4$, estimated using DES/*Euclid* photometry. We annotate this panel with the estimated $\Delta_{\langle z\rangle}$ made when utilising our two different priors. It is clear that the choice of prior, in this circumstance, can have a significant impact on the recovered redshift distribution. We also find an offset in the estimated redshift distributions with respect to the truth, as confirmed by the associated mean redshift biases being considerable, $|\Delta_{\langle z\rangle}| > 0.012$, which is roughly six times larger than the *Euclid* accuracy requirement.

The resulting biases estimated for this method in all tomographic bins, averaged over all noise realisations, is presented in the left-most panels of Fig. 5 (for both the DES/*Euclid* and LSST/*Euclid* configurations). Overall, we find that this approach produces mean biases of $|\mu_{\Delta z}| > 0.02\,(1+z)$ and $|\mu_{\Delta z}| > 0.01\,(1+z)$, which correspond to roughly ten and five times larger than the *Euclid* accuracy requirement for the DES/*Euclid* and LSST/*Euclid* cases, respectively. Such bias is created by the mismatch between the simple galaxy templates included in `LePhare` (in a broad sense, including dust attenuation and intergalactic medium absorption) and the complexity and diversity of galaxy spectra generated in the hydrodynamical simulation. Such biases are in agreement with the usual values observed in the literature with broadband data (e.g., Hildebrandt et al. 2012). We therefore conclude that the use of such a redshift calibration method is not feasible for *Euclid*, even under optimistic photometric circumstances.

4.3. Redshift probability de-biasing

In the previous section, we demonstrated that the estimation of galaxy redshift distributions via the summation of individual galaxy posteriors, $\mathcal{P}(z)$, estimated with a standard template-fitting code, is too inaccurate for the requirements of the *Euclid* survey. The cause of this inaccuracy can be traced to a number of origins: colour-redshift degeneracies, template set non-representativeness, redshift prior inadequacy, and more. However, it is possible to alleviate some of this bias, statistically, by incorporating additional information from a spectroscopic training sample. In particular, Bordoloi et al. (2010) proposed a method to de-bias $\mathcal{P}(z)$ distributions using the probability integral transform (PIT, Dawid 1984). The PIT of a distribution is defined as the value of the cumulative distribution function evaluated at the ground truth. In the case of redshift calibration, the PIT per galaxy is therefore the value

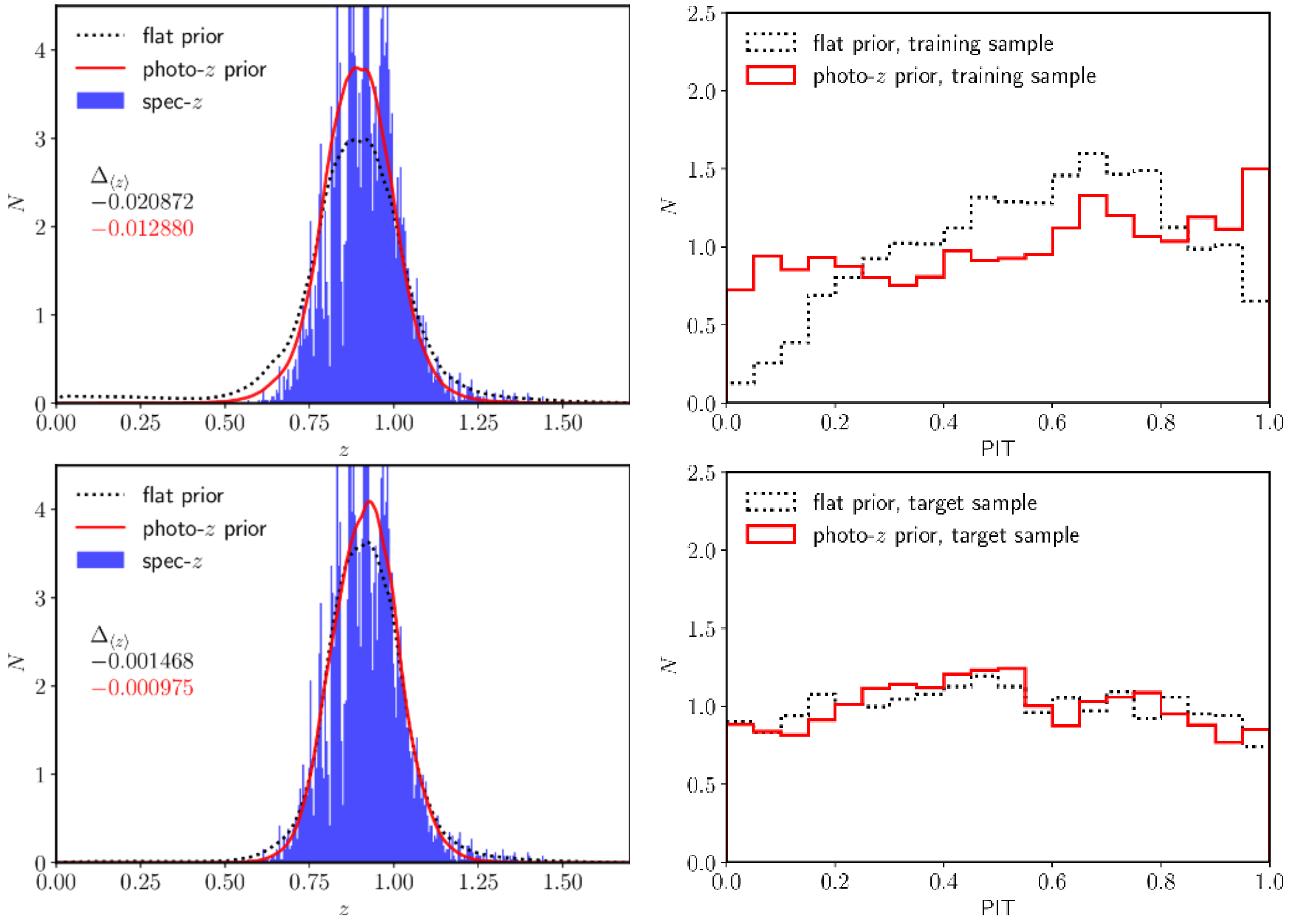

Fig. 4. Examples of redshift distributions (*left*) and PIT distributions (*right*; see text for details) for a tomographic bin selected to $0.8 < z_\mathrm{p} < 1$ using DES/*Euclid* photo-z. In these examples, we assume a training sample extracted from a SOM, with two galaxies per cell. *Top* and *bottom panels*: results before and after zPDF de-biasing, respectively. Redshift distributions and PITs are shown for the true redshift distribution (blue) and redshift distributions estimated using the zPDF method when incorporating photo-z (red) and uniform (black) priors.

of the cumulative $\mathcal{P}(z)$ distribution evaluated at source spectroscopic redshift z_s:

$$\mathrm{PIT} = C(z_\mathrm{s}) = \int_0^{z_\mathrm{s}} \mathcal{P}(z)\,\mathrm{d}z. \tag{13}$$

If all the individual galaxy redshift probability distributions are accurate, the PIT values for all galaxies should be uniformly distributed between 0 and 1. Therefore, using a spectroscopic training sample, any deviation from uniformity in the PIT distribution can be interpreted as an indication of bias in individual estimates of $\mathcal{P}(z)$ per galaxy. We define N_P as the PIT distribution for all the galaxies within the training spectroscopic sample in a given tomographic bin. Bordoloi et al. (2010) demonstrate that the individual $\mathcal{P}(z)$ can be de-biased using the N_P as

$$\mathcal{P}_\mathrm{deb}(z) = \mathcal{P}(z) \times N_\mathrm{P}[C(z)] \left[\int_0^1 N_\mathrm{P}(x)\,\mathrm{d}x\right]^{-1}, \tag{14}$$

where $\mathcal{P}_\mathrm{deb}(z)$ is the de-biased posterior probability and the last term ensures correct normalisation. This correction is performed per tomographic bin.

This method assumes that the correction derived from the training sample can be applied to all galaxies of the target sample. As with the direct calibration method, such an assumption is valid only if the training sample is representative of the target sample (i.e. in the case of a uniform training sample), which is not the case for the COSMOS-like or SOM training samples. In these cases, we weight each galaxy of the training sample in a manner equivalent to the direct calibration method (see Sect. 3) in order to ensure that the PIT distribution of the training sample matches that of the target sample (which is, of course, unknown). As for direct calibration, a completely missing population (in redshift or spectral type) could impact the results in an unknown manner, but such a case should not occur for a uniform or SOM training sample.

Until now, we have considered two types of redshift prior (defined in Sect. 4.2): (1) the flat prior and (2) the photo-z prior. We have shown that the choice of prior can have a significant impact on the recovered $\langle z\rangle$ (Sect. 4.2). However, as already noted by Bordoloi et al. (2010), the PIT correction has the potential to account for the redshift prior implicitly. In particular, if one uses a flat redshift prior, the correction essentially modifies $\mathcal{L}(z)$ to match the true $\mathcal{P}(z)$ (if the various abovementioned assumptions are satisfied). This is because the redshift prior information is already contained within the training spectroscopic sample. Nonetheless, rather than assuming a flat prior to measure the PIT distribution, one can also adopt the photo-z prior (as in Eq. (12)). This approach has two advantages: (1) It allows us to start with a posterior probability that is intrinsically closer to the truth, and (2) it includes the magnitude dependence of the redshift distribution within the prior, which is, of course, not reflected in the case of the flat prior.

Therefore, we improved the de-biasing procedure from Bordoloi et al. (2010) by including such a photo-z prior. We added an iterative process to further ensure the correction's fidelity and stability. In this process, the PIT distribution is iteratively recomputed by updating the photo-z prior. We computed the PIT for the galaxy as

$$C^n(z_s) = \int_0^{z_s} \mathcal{L}(z)\,\mathrm{Pr}^n(z|m_0)\,\mathrm{d}z, \tag{15}$$

where $\mathrm{Pr}^n(z|m_0)$ is the prior computed at step n. We can then derive the de-biased posterior as

$$\mathcal{P}^n_{\mathrm{deb}}(z) = \mathcal{L}(z)\,\mathrm{Pr}^n(z|m_0) \times N^n_{\mathrm{P}}[C^n(z)], \tag{16}$$

where N^n_{P} is the PIT distribution at step n. The prior at the next step is

$$\mathrm{Pr}^{n+1}(z|m_0) = \sum_i^{N_\mathrm{T}} \mathcal{P}^n_{\mathrm{deb},i}(z|\boldsymbol{o})\,\Theta(m_i|m_0), \tag{17}$$

where m_i is the magnitude of the galaxy i. It should be noted that we assume a flat prior at $n = 0$. Therefore, the step $n = 0$ of the iteration corresponds to the de-biasing assuming a flat prior, as in Bordoloi et al. (2010). We also note that the prior is computed for the N_T galaxies of the training sample in the de-biasing procedure, while it is computed over all galaxies of the tomographic bin for the final posterior.

As an illustration, Fig. 2 shows the de-biased posterior distributions with black lines, which can significantly differ from the original likelihood distribution. We find that this procedure converges quickly. Typically, the difference between the mean redshift measured at step $n + 1$ and that measured at step n does not differ by more than 10^{-3} after two to three iterations.

As described in Appendix A, we also find that the de-biasing procedure is considerably more accurate when the photo-z uncertainties are overestimated, rather than underestimated. Such a condition can be enforced for all galaxies by artificially inflating the source photometric uncertainties by a constant factor in the input catalogue prior to the measurement of photo-z. In our analysis, we utilised a factor of two inflation in our photometric uncertainties prior to the measurement of our photo-z in our de-biasing technique.

4.4. Final results

We illustrate the impact of the $\mathcal{P}(z)$ de-biasing on the recovered redshift distribution in the lower panels of Fig. 4. This figure presents the case of the redshift bin $0.8 < z_\mathrm{p} < 1$ in the DES/*Euclid* configuration. The $N(z)$ and PIT distributions, as computed with the initial posterior distribution, are shown in the upper panels (for both of our assumed priors). The distributions after de-biasing are shown in the bottom panels. We can see the clear improvement provided by the de-biasing procedure in this example, whereby the redshift distribution bias $\Delta_{\langle z \rangle}$ (annotated) is reduced by a factor of ten. We also observe a clear flattening of the target sample PIT distribution.

We present the results of de-biasing on the mean redshift estimation for all tomographic bins in Fig. 5. The three rightmost panels show the mean redshift biases recovered by our de-biasing method, averaged over the 18 photometric noise realisations, for our three training samples. The accuracy of the mean redshift recovery is systematically improved compared to the case without $\mathcal{P}(z)$ de-biasing (shown in the left column). In the DES/*Euclid* configuration, for instance (shown in the upper row), the improvement is better than a factor of ten at $z > 1$. In the LSST/*Euclid* configuration (shown in the bottom row), we find that the results do not depend strongly on the training set used: The accuracy of $\langle z \rangle$ is similar for the three training samples, showing that stringent control of the representativeness of the training sample is not necessary in this case. In the DES/*Euclid* case, however, the SOM training sample clearly outperforms the other training samples, especially at low redshifts. Finally, we note that the iterative procedure using the photo-z prior improves the results when using the SOM training sample and the DES/*Euclid* configuration.

Overall, the *Euclid* requirement on redshift calibration accuracy is not reached by our de-biasing calibration method in the DES/*Euclid* configuration. The values of $\mu_{\Delta z}$ at $z < 1$ are five times too high compared to the *Euclid* requirement, represented by the yellow bands in Fig. 5. At best, an accuracy of $|\mu_{\Delta z}| \leq 0.004\,(1+z)$ is reached for the SOM training sample with the photo-z prior. Conversely, the *Euclid* requirement is largely satisfied in the LSST/*Euclid* configuration. In this case, biases of $|\mu_{\Delta z}| \leq 0.002\,(1+z)$ are observed in all but the two most extreme tomographic bins: $0.2 < z < 0.4$ and $2 < z < 2.2$. We therefore conclude that, for this approach, deep imaging data are crucial for reaching the required accuracy on mean redshift estimates for *Euclid*.

5. Discussion on key model assumptions

In this section, we discuss how some important parameters or assumptions impact our results. We start by discussing the impact of catastrophic redshift failures in the training sample, the impact of our preselection on photometric redshift uncertainty, and the influence of the size of the training sample on our conclusions. We also discuss some remaining limitations of our simulation in the last subsection.

5.1. Impact of catastrophic redshift failures in the training sample

For all results presented in this work so far, we have assumed that spectroscopic redshifts perfectly recover the true redshift of all training sample sources. However, given the stringent limit on the mean redshift accuracy in *Euclid*, deviations from this assumption may introduce significant biases. In particular, mean redshift estimates are extremely sensitive to redshifts far from the main mode of the distribution, and therefore catastrophic redshift failures in spectroscopy may present a particularly significant problem. For instance, if 0.5% of a galaxy population with a true redshift of $z = 1$ are erroneously assigned $z_s > 2$, then this population will exhibit a mean redshift bias of $|\mu_{\Delta z}| > 0.002$ under direct calibration.

Studies of duplicated spectroscopic observations in deep surveys have shown that there exists, typically, a few percent of sources that are assigned both erroneous redshifts and high confidences (e.g., Le Fèvre et al. 2005). Such redshift measurement failures can be due to misidentification between emission lines, incorrect associations between spectra and sources in photometric catalogues, and/or incorrect associations between spectral features and galaxies (due, for example, to the blending of galaxy spectra along the line of sight Masters et al. 2017; Urrutia et al. 2019). Of course, the fraction of redshift measurement failures is dependant on the observational strategy (e.g., spectral resolution) and the measurement technique (e.g., the

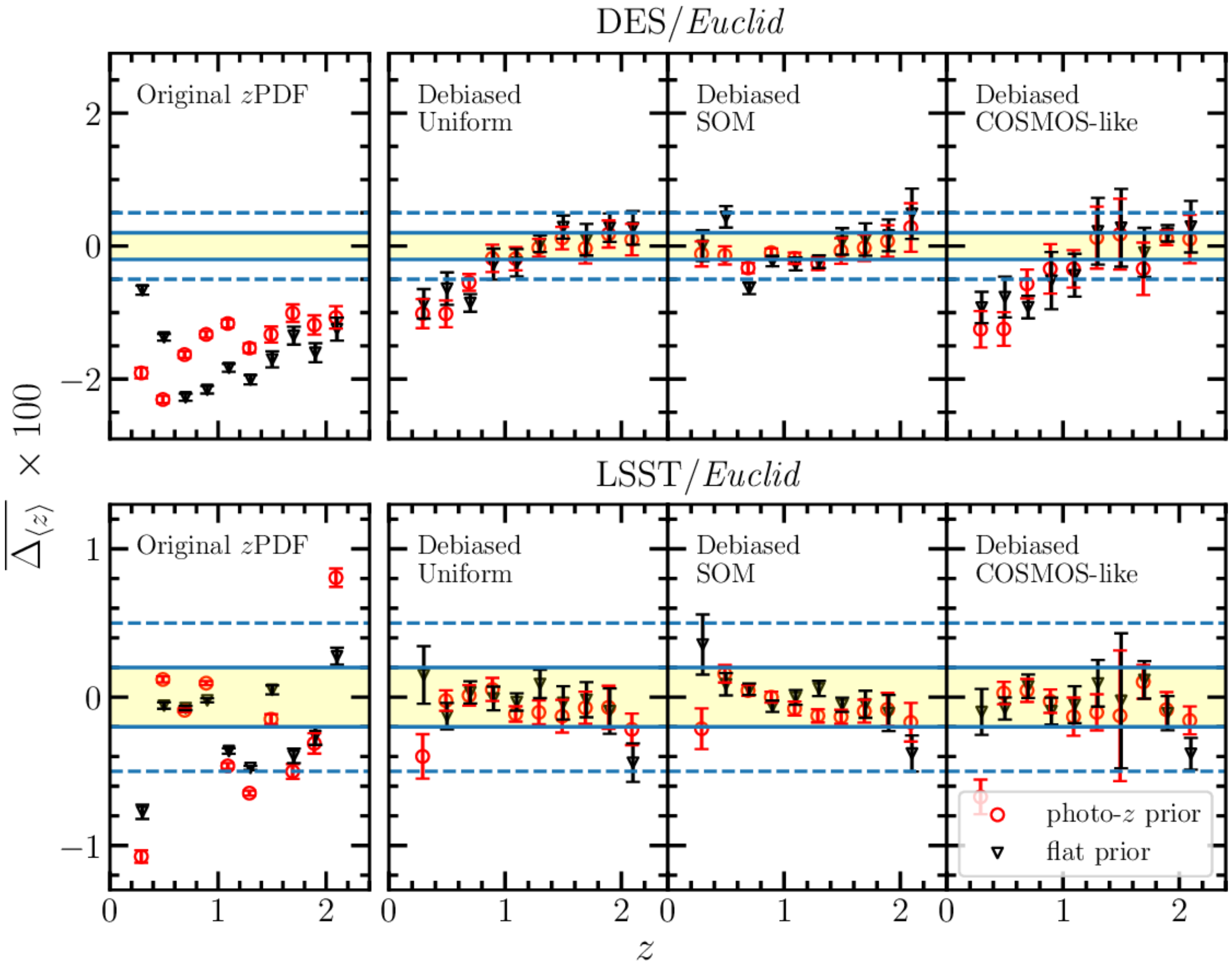

Fig. 5. Bias on the mean redshift (see Eq. (3)) estimated using the zPDF method and averaged over the 18 photometric noise realisations. *Top* and *bottom panels*: correspond to the mock DES/*Euclid* and LSST/*Euclid* catalogues, respectively. We note the differing scales in the y-axes of the two panels. *Left panels*: are obtained by summing the initial zPDF without any attempt at de-biasing. The other panels show the results of summing the zPDF after de-biasing, assuming (*from left to right*) a uniform, SOM, and COSMOS-like training sample. The yellow region represents the *Euclid* requirement of $|\Delta_{\langle z\rangle}| \leq 0.002\,(1+z)$. The red circles and black triangles in each panel correspond to the results estimated using photo-z and flat priors, respectively.

number of reviewers per observed spectrum). The incorrect association of stars and galaxies can also create difficulties. Furthermore, the frequency of redshift measurement failures is expected to increase as a function of source apparent magnitude, which is a particular problem for the faint sources probed by *Euclid* imaging (VIS < 24.5).

As we cannot know a priori the number (nor location) of catastrophic redshift failures in a real spectroscopic training set, we instead estimated the sensitivity of our results to a range of catastrophic failure fractions and modes. We assumed a SOM-based training sample and an LSST/*Euclid* photometric configuration and distributed various fractions of spectroscopic failures throughout the training sample, simulating both random and systematic failures. Generally, though, because these failures occur in the spectroscopic space, recovered calibration biases are largely independent of the depth of the imaging survey and the method used to build the training sample.

We started by testing the simplest possible mechanism of distributing the failed redshifts, by assigning failed redshifts uniformly within the interval $0 < z < 4$. Resulting calibration biases for this mode of catastrophic redshift failure are presented in the left panels of Fig. 6. We find that, for the direct calibration approach (top panel), the limit to bias the mean redshift by $|\mu_{\Delta z}| > 0.002$ at low redshifts in the training sample is as low as 0.2% of failures (by definition, flag 3 in the VIMOS VLT Deep Survey (VVDS) could include 3% of failures; Le Fèvre et al. 2005). We also find that the bias decreases with redshift and reaches zero at $z = 2$. This is a statistical effect; our assumed uniform distribution has a $z = 2$ mean, and so random catastrophic failures scattered about this point induce no shift in a $z \approx 2$ tomographic bin. For the same reason, biases would be significant in the two extreme tomographic bins if we were to assume a catastrophic failure distribution that followed the true $N(z)$ (which peaks at $z \approx 1$). In contrast, our de-biased zPDF approach is found to be resilient to catastrophic failure fractions as high as 3.0% (bottom panel). In that case, only an unlikely failure fraction of 10% would bias the mean redshift by $|\mu_{\Delta z}| \geq 0.002\,(1+z)$. We interpret this result as a demonstration of the low sensitivity

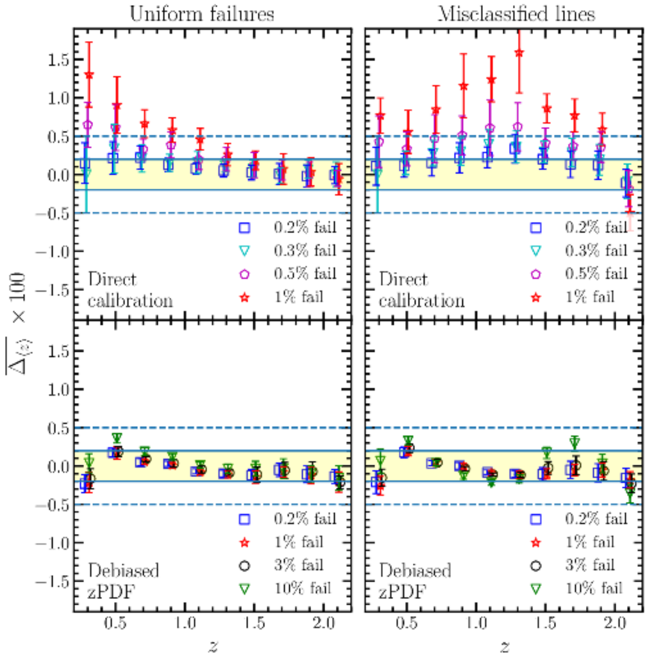

Fig. 6. Bias on the mean redshift averaged over the 18 photometric noise realisations in the LSST/*Euclid* case. We assume a SOM training sample, and the different symbols correspond to various fractions of failure introduced in the spec-z training sample. *Left* and *right panels*: correspond to different assumptions on how to distribute the catastrophic failures in the spec-z measurements: uniformly distributed between $0 < z < 4$ (*left*) and assuming the failures are caused by misclassified emission lines (*right*). *Upper* and *lower panels*: correspond to the direct calibration and de-biasing methods, respectively.

of the PIT distribution to redshift failures in the training sample. This is related to the fact that the PIT distribution provides a global statistical correction that is only weakly sensitive to individual galaxy redshifts.

In the previous test, we assigned the failed redshifts uniformly within the interval $0 < z < 4$, which is not the expected distribution when redshift failures occur from the misidentification of spectral emission lines (e.g., Le Fèvre et al. 2015; Urrutia et al. 2019). This mode of failure leads to a highly non-uniform distribution of failed redshifts due to the interplay between the location of spectral emission lines and the redshift distribution of training sample galaxies. If a line emitted at $\lambda_{\rm true}$ is misclassified as a different emission line at $\lambda_{\rm wrong}$, the redshift is therefore assigned to be

$$z_{\rm wrong} = \frac{\lambda_{\rm true}}{\lambda_{\rm wrong}}(1 + z_{\rm true}) - 1. \quad (18)$$

We studied the impact of such line misidentifications on our estimates of $\langle z\rangle$ by introducing redshift failures in the simulation with the following four assumptions: (1) If $z_{\rm true} < 0.5$, we assume that the Hα emission line can be misclassified as [OII]; (2) if $0.5 < z_{\rm true} < 1.4$, we assume that [OII] can be misclassified as Hα (for bright sources) or Lyα (for faint sources, using $i = 23.5$ as a limit); (3) at $1.4 < z_{\rm true} < 2.0$, we assume that the redshift is estimated using NIR spectra and therefore that the Hα line can be misclassified as [OII]; and (4) for sources at $z > 2$, we assume that Lyα can be misclassified as [OII].

The same fraction of misclassifications is assumed in all the redshift intervals. The result of this experiment is shown in the right panels of Fig. 6 and demonstrates that this (more realistic) mode of catastrophic failures results in equivalent levels of bias as was seen in our simple (uniform) mode, albeit in different tomographic bins. This confirms that the sensitivity of the direct calibration method to catastrophic redshift failures exists across simplistic and complex failure modes. In this mode, a failure fraction of 0.2% is sufficient to bias direct calibration at $|\mu_{\Delta z}| \geq 0.002\,(1 + z)$ in all tomographic bins with $z_{\rm p} > 0.6$. This highlights that the calibration bias depends on the exact distribution of failed redshifts: In the case of line misidentification, incorrectly assigned redshifts consistently bias spectra to higher redshifts, causing $\langle z\rangle$ to be affected more heavily over the full redshift range.

We compared our result to the simulation of Wright et al. (2020). They investigate the impact of catastrophic spec-z failures on the estimate of $\langle z\rangle$ (for KiDS cosmic shear analyses) in the MICE2 simulation (Fosalba et al. 2015). They introduced 1.03% of failed redshifts following various distributions. In particular, they tested the case of a uniform distribution within $0 < z < 1.4$, where $z = 1.4$ is the limiting redshift of the MICE2 simulation. They report a bias in their direct calibration of $\Delta_{\langle z\rangle} = 0.0029$ for their lowest redshift tomographic bin, and smaller biases for higher redshift tomographic bins. In our lowest redshift bin, we observe a bias of $\Delta_{\langle z\rangle} = 0.01$ for a similar analysis. We argue that this is entirely consistent with the results of Wright et al. (2020) given that our considered redshift range is almost three times larger. Wright et al. (2020) conclude that spec-z failures are unlikely to influence cosmic shear analyses with the KiDS survey, which are limited to $z < 1.2$, but may be significant for *Euclid*-like analyses. In this way, our results also agree; it is clear that direct calibration for next generation (so-called Stage IV) cosmic-shear surveys such as *Euclid* will require careful consideration of the influence of catastrophic spectroscopic failures.

The training sample for *Euclid* is currently being built with the C3R2 survey (Masters et al. 2019; Guglielmo et al. 2020). Such a sample results from a combination of spectra coming from numerous instruments installed on 8-metre class telescopes (e.g., VIMOS, FORS2, KMOS, DEIMOS, LRIS, and MOSFIRE) and including data from previous spectroscopic surveys (e.g., Lilly et al. 2007; Le Fèvre et al. 2015; Kashino et al. 2019). The most robust spec-z acquired on the *Euclid* Deep Fields with the NISP instrument will be included. Given the diversity of observations, a careful assessment of the sample purity is necessary to limit the fraction of failures below 0.2%. Encouragingly, Masters et al. (2019) do not find any redshift failures within the 72 C3R2 spec-z with duplicated observations. Nonetheless, a larger sample of confirmed spectra is necessary to demonstrate that fewer than 0.2% of spectroscopic redshift measurements suffer catastrophic failure. Finally, it is possible that the improved reliability of both direct calibration methods and spectroscopic confidence could decrease the effects seen here: Wright et al. (2020), for example, advocate a means of cleaning cosmic shear photometric samples of sources with poorly constrained mean redshifts, demonstrating that this can cause a considerable reduction in calibration biases. Of course, the problem could possibly be alleviated if one were able to improve the reliability of the training sample by only including spec-z with corroborative evidence from, for example, high-precision photo-z derived from deep photometry in the calibration fields.

5.2. Relaxing the photo-z σ_{z_p} preselection

Estimates of the redshift distribution mean are also sensitive to the presence of secondary modes in the redshift distribution, as well as our ability to reconstruct them. As described in Sect. 2.2,

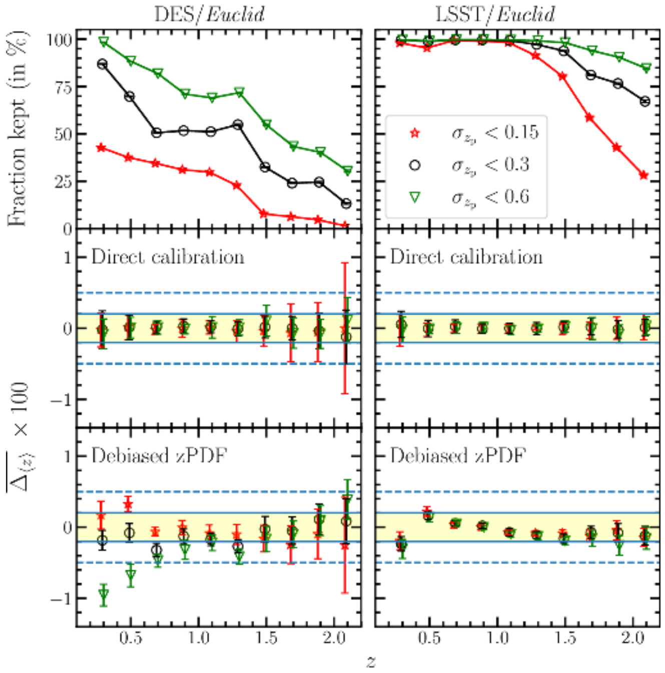

Fig. 7. Bias on the mean redshift (see Eq. (3)), averaged over the 18 photometric noise realisations, under different σ_{z_p} selection thresholds. *Top panels*: fraction of the sample retained after having applied different σ_{z_p} thresholds. *Middle* and *bottom panels*: bias on the mean redshift using the direct calibration and de-biasing techniques, respectively. The left and right panels correspond to the DES/*Euclid* and LSST/*Euclid* configurations, respectively. We assume a SOM training sample with 2 galaxies per cell.

all results presented thus far have invoked a selection on the photometric redshift uncertainty of $\sigma_{z_p} < 0.3$, which reduces the likelihood of secondary redshift distribution peaks in our analysis. Here we discuss the impact of this adopted threshold on both the accuracy of our estimates of $\langle z \rangle$ and on the fraction of photometric sources that satisfy this selection (and so are retained for subsequent cosmic shear analysis). We applied several σ_{z_p} thresholds in the range $\sigma_{z_p} \in [0.15, 0.6]$ to the full photo-z catalogue. For the training sample, we considered the SOM configuration with two galaxies per cell. The results are shown in Fig. 7 for the DES/*Euclid* (left) and LSST/*Euclid* (right) configurations. We find that the σ_{z_p} threshold does not influence our conclusions regarding the direct calibration approach, which is largely insensitive to variations in this threshold. We note, however, that the scatter on the mean redshift ($\sigma_{\Delta z}$, shown by the errorbars) increases well above the *Euclid* requirement (for the DES/*Euclid* configuration) when selecting photo-z with $\sigma_{z_p} < 0.15$; however, this is primarily because such a selection drastically reduces the size of the training sample at $z > 1.2$, increasing the influence of Poisson noise. Therefore, given the insensitivity of the direct calibration to this threshold, it is advantageous to keep galaxies with broad redshift likelihoods in the target sample when using this method. Conversely, σ_{z_p} has a decisive impact on the accuracy of mean redshift estimates inferred from the de-biased zPDF approach. For instance, in the DES/*Euclid* configuration, $|\mu_{\Delta z}|$ is strongly degraded when applying a threshold of $\sigma_{z_p} < 0.6$. Such a threshold on σ_{z_p} could be relaxed in the LSST/*Euclid* configuration, however, primarily because the sample is already dominated by galaxies with a narrow zPDF.

Not considered in the above, however, is the importance that the target sample number density plays in cosmic shear analyses. Cosmological constraints from cosmic shear are approximately proportional to the square root of the size of the target galaxy sample, as well as to the mean redshift. Therefore, optimal lensing surveys require a sufficiently high surface density of sources, preferentially at high redshifts. In the *Euclid* project, 30 galaxies per arcmin^2 are required to reach their planned scientific objectives (Laureijs et al. 2011). As shown in the top panels of Fig. 7, however, applying a threshold on σ_{z_p} naturally introduces a reduction in the size of the target sample. For instance, we keep fewer than 10% of the galaxies at $z > 1.4$ by selecting a sample at $\sigma_{z_p} < 0.15$ in the DES/*Euclid* configuration. In the LSST/*Euclid* case, a threshold of $\sigma_{z_p} < 0.3$ only has a significant impact in the redshift bins above $z > 1.6$. A compromise is therefore needed between the number of sources retained in the target sample and the accuracy of the mean redshift that we estimate for these sources (when using the de-biasing technique). We have not attempted to estimate what this optimal selection would be using our simulations as the luminosity function predicted by Horizon-AGN does not perfectly reproduce what is found in real data. Nonetheless, we note that the fraction of galaxies that are removed from the target sample is likely overestimated here: Modern cosmic shear analyses typically introduce a weight associated with the accuracy of each source's shape measurement (the 'shear weight', which is not included in our simulations), which systematically decreases the contribution of low signal-to-noise galaxies to the analysis. As these fainter sources have intrinsically broader photo-z distributions, they will be the most heavily affected by our cuts on σ_{z_p}.

5.3. Size of the training sample

The size of the training sample is naturally of the highest importance when using the direct calibration approach (e.g., Newman 2008). The de-biased zPDF approach, though, is also sensitive to statistical noise in the PIT distribution. As some ongoing spectroscopic surveys are designed to produce the training samples for Stage IV weak-lensing experiments (e.g., Masters et al. 2017), we explore here the minimal size of these samples required for accurate redshift calibration. To do this, we modified the size of the training samples (limiting our analysis to the uniform and SOM training sample cases). We did not consider the COSMOS-like case that is a patchwork of existing surveys and which is not specifically designed for weak-lensing experiments. For the uniform training samples, we tested the cases with 500, 1000, and 2000 galaxies per tomographic bin. For the SOM training samples, we tested the cases corresponding to cells filled with one, two, or three galaxies.

Figure 8 shows the impact of the training sample size on $\Delta_{\langle z \rangle}$. We find that the mean bias $\mu_{\Delta z}$ always remains within the *Euclid* requirements for the direct calibration approach. The scatter $\sigma_{\Delta z}$ in the bias exceeds the *Euclid* requirements in a few tomographic bins, though only when considering the smallest training samples: The *Euclid* requirements are fully satisfied in all tomographic bins when assuming a training sample with more than 1000 galaxies per bin or more than two galaxies per SOM cell. With the de-biased zPDF approach, we find that increasing the size of the training sample is not sufficient to reduce the residual bias in the method; instead, deeper photometry is preferable for improving the quality of the initial zPDF.

5.4. Catastrophic failures within the photo-z sample

Catastrophic failures in the photo-z sample are a concern for both of the methods described in this paper. We discuss

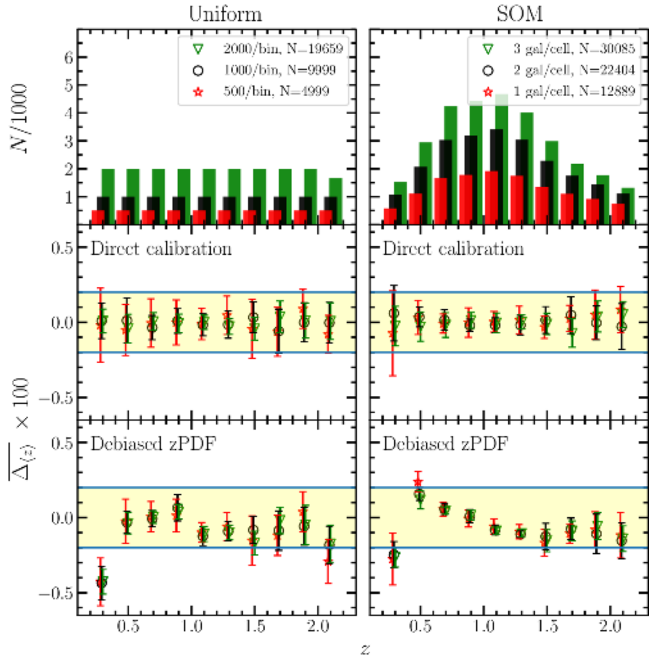

Fig. 8. Bias on the mean redshift (see Eq. (3)) averaged over the 18 photometric noise realisations and the impact of the training sample size on the mean redshift accuracy in the LSST/*Euclid* case. *Left* and *right panels*: correspond to uniform and SOM spectroscopic coverage, respectively. *Top panels*: number of galaxies used for the training in the three considered cases. *Middle* and *bottom panels*: mean redshift accuracy using the direct calibration and the optimised zPDF methods, respectively.

here their impact as well as the remaining limitations of our simulation.

As shown in Fig. 1, our simulated sample already includes a significant fraction of photo-z outliers, defined such that $|z_\mathrm{p} - z_\mathrm{s}| > 0.15\,(1 + z_\mathrm{s})$. We find 16.24% and 0.70% of outliers at VIS < 24.5 in DES/*Euclid* and LSST/*Euclid*, respectively. These fractions reduce to 1.82% and 0.04% when applying a selection on the photometric redshift uncertainty at $\sigma_{z_\mathrm{p}} < 0.3$. The largest fraction of these outliers is due to the degeneracies in the colour-redshift space inherent to the use of low signal-to-noise photometry in several bands. However, less trivial catastrophic failures are also present in the simulation. In particular, the diversity of spectra generated by the complex physical processes in Horizon-AGN is not fully captured by the limited set of SED templates used in `LePhare`. This misrepresentation in galaxy SED creates a significant fraction of zPDFs that are not compatible with the spec-z. An example of such an $\mathcal{L}(z)$ is shown in the bottom-right panel of Fig. 2. Despite the presence of such failures, our results show that the *Euclid* requirement is fulfilled.

Several factors that could potentially create more catastrophic failures in the photo-z were ignored. Galaxies with extreme properties, such as sub-millimetre galaxies (SMGs), are known to be underrepresented in simulations (e.g., Hayward et al. 2021). If galaxies with an extreme dust attenuation fall within the cosmic-shear selection at VIS < 24.5 and are selected in one tomographic bin, they could have an impact on our results. Nonetheless, nothing indicates that their zPDF cannot be correctly established from template fitting, nor that such a population cannot be isolated in the multi-colour space with a SOM.

The presence of AGN could also be a problem. These sources can be isolated from their SEDs (Fotopoulou & Paltani 2018), identified as point-like sources for quasi-stellar objects, and identified as X-ray sources with eROSITA (Merloni et al. 2012). We should, though, fail to isolate AGN with extended morphologies or that are too faint to be detected in X-ray. Salvato et al. (2011) find, however, that standard galaxy SED libraries are sufficient to obtain accurate photo-z for such sources.

Residual contamination from stars could also bias $\langle z\rangle$. This population preferentially contaminates specific tomographic bins. In particular, stars may bias the mean redshift towards higher values for both the direct calibration and de-biased zPDF methods. A morphological selection based on high-resolution VIS images, combined with a colour selection that includes NIR photometry (e.g., Daddi et al. 2004), is efficient at isolating them (Fotopoulou & Paltani 2018). A minimal contamination could bias the mean redshift at a level similar to the one discussed in Sect. 5.1. Nonetheless, future simulations need to include stellar and AGN populations to better assess the level of contamination of the galaxy sample and its impact on the *Euclid* requirement.

Finally, Laigle et al. (2019) show that the fraction of outliers in Horizon-AGN remains underestimated relative to the real dataset. One source of discrepancy originates from not taking the uncertainties induced by source extraction in images into account. Bordoloi et al. (2010) estimate that 10% of the sources could potentially be blended and that the likelihood of two blended galaxies with a magnitude difference lower than two is affected in an unpredictable way. Over the last decade, numerous source extraction methods have been developed to perform photometry in crowded fields (De Santis et al. 2007; Laidler et al. 2007; Merlin et al. 2016; Lang et al. 2016), which could mitigate the impact of blending. Therefore, a new set of simulations that include images and such source extraction tools should be considered in the future.

6. Application to real data

In this section, we apply the two approaches presented in Sects. 3 and 4 to real data. We use existing imaging surveys and associated photo-z to define several tomographic bins. In each tomographic bin, we select a sub-sample of spec-z for which the mean redshift $\langle z\rangle_\mathrm{true}$ is known. We refer to this sample as the target sample, and the goal is to retrieve the mean redshift using only the photometric catalogue and an independent training sample. As previously, we measure $\Delta_{\langle z\rangle}$ as defined in Eq. (3) in each tomographic bin.

6.1. The COSMOS survey

We first investigated a favourable configuration, where the photometric survey is much deeper than the target sample. We aim at measuring the mean redshift of the Large Early Galaxy Astrophysics Census (LEGA-C) galaxies (van der Wel et al. 2016) selected in the tomographic bin at $0.7 < z_\mathrm{p} < 0.9$. We based our estimate of $\langle z\rangle$ on the COSMOS broadband photometry and associated zPDF. The imaging sensitivity is three magnitudes deeper than that of the target sample. All the spec-z available on the COSMOS field (excluding the LEGA-C ones) are used for the training. For the direct calibration approach, we obtain a bias of $\mu_{\Delta z} = 0.00032$ and a scatter of $\sigma_{\Delta z} = 0.00135$, an accuracy well within the *Euclid* requirement. Secondly, we de-biased the zPDF using the PIT distribution as discussed in Sect. 4.3. In that case, we obtain a mean redshift with a bias of $\mu_{\Delta z} = -0.00046$ and a scatter of $\sigma_{\Delta z} = 0.00073$. In the case of a target sample associated with much deeper photometry, we thus reach the $0.002\,(1 + z)$ accuracy requirement of *Euclid*, using either the

Table 1. Differences between the mean redshifts reconstructed with different methods (direct calibration and de-biased zPDF) and $\langle z\rangle_{\rm true}$, divided by $(1 + \langle z\rangle_{\rm true})$.

$z_{\rm min}$	$z_{\rm max}$	% kept	$N_{\rm train}$	Direct calib. $[10^{-2}]$	zPDF *w/* flat prior $[10^{-2}]$	zPDF *w/* photo-z prior $[10^{-2}]$
				$\sigma_{z_{\rm p}} < 0.3$		
0.10	0.30	79.80	1192.00	1.72	2.78	0.94
0.30	0.50	72.10	2156.00	0.64	0.33	0.36
0.50	0.70	55.60	1497.00	−0.57	−0.88	−0.28
0.70	0.90	68.70	1822.00	−0.65	−1.38	−0.89
0.90	1.20	62.00	892.00	0.10	0.29	−0.22
				$\sigma_{z_{\rm p}} < 0.6$		
0.10	0.30	96.60	1318.00	1.34	3.19	−0.88
0.30	0.50	89.40	2321.00	−0.56	0.48	−0.40
0.50	0.70	80.80	1845.00	−1.26	−2.60	−1.50
0.70	0.90	89.60	2094.00	−0.34	−1.75	−0.79
0.90	1.20	81.70	1057.00	0.38	1.16	−0.03
				$\sigma_{z_{\rm p}} < 1.2$		
0.10	0.30	97.80	1326.00	1.37	3.50	−1.01
0.30	0.50	93.90	2357.00	−0.38	0.90	−0.46
0.50	0.70	88.20	1886.00	−0.92	−2.42	−1.63
0.70	0.90	93.70	2131.00	−0.11	−1.67	−0.92
0.90	1.20	90.40	1116.00	1.66	2.67	0.43

Notes. The KiDS+VIKING-450 survey is split into five tomographic bins. We use VVDS/DEEP2 as the target sample and COSMOS as the training one. In the top part of the table, photo-z are selected with $\sigma_{z_{\rm p}} < 0.3$, while the bottom parts show a selection at $\sigma_{z_{\rm p}} < 0.6$ and $\sigma_{z_{\rm p}} < 1.2$. The fraction of galaxies kept after this selection is also shown ('% kept'). We apply the same definition as Wright et al. (2020) to define the loss of photometric sources (their Eq. (1)), including shear weights.

direct calibration or de-biased zPDF approaches. The details of this measurement are given in Appendix B.

6.2. The KiDS+VIKING-450 survey

We now study a less favourable case where the photometric survey has a similar depth as the target sample. We measured the mean redshift in five tomographic bins extracted from the KiDS+VIKING-450 imaging survey, which covers 341 deg^2 (Wright et al. 2019). The survey combines the *ugri*-band photometry from KiDS with the $ZYJHK_s$ bands from VISTA Kilo degree Infrared Galaxy (VIKING) photometry. We adopted the method described in Sect. 2.2 to measure the photo-z. This leads to a photo-z quality comparable to that obtained by Wright et al. (2019), where $\sigma_{\rm NMAD} \sim 0.045$ at $z < 0.9$ and $\sigma_{\rm NMAD} \sim 0.079$ at $z > 0.9$. These photo-z were used to define five tomographic bins over the photometric redshift interval $0.1 < z < 1.2$, as in Hildebrandt et al. (2020).

The KiDS+VIKING-450 survey encompasses the VVDS (Le Fèvre et al. 2005) and DEEP2 (Newman et al. 2013) fields, which contain spectroscopic redshifts. We aim at retrieving the mean redshift of the VVDS/DEEP2 galaxies. By only selecting galaxies with secure spectroscopic redshifts and counterparts in the KiDS+VIKING-450 catalogue, we built a target sample of 5794 galaxies[3]. The DEEP2 sample was selected at $R < 24.1$ and $z > 0.7$, while the VVDS sample was purely magnitude-limited at $i < 24$. Our target sample covers the full redshift range of interest $0.1 < z < 1.2$, with magnitude limits similar to those used for the KiDS+VIKING-450 cosmic shear analysis (Hildebrandt et al. 2020).

The KiDS+VIKING-450 imaging survey also covers the COSMOS field, and we used the existing spec-z in the COSMOS field as the training sample. We note that the training and target samples are located in different fields. Therefore, the sample variance may impact our results. The COSMOS training sample contains 13 817 galaxies from the KiDS+VIKING-450 survey, after applying a redshift confidence selection. This highly heterogeneous sample combines various spectroscopic surveys covering a large range of magnitudes and redshifts (see Sect. 2.3 and Laigle et al. 2016, for more details).

We present our results in Table 1 for the five considered tomographic bins. The upper section of the table shows the fiducial case, where a $\sigma_{z_{\rm p}} < 0.3$ photo-z uncertainty selection is applied. The direct calibration produces a bias of $|\Delta_{\langle z\rangle}| < 0.01\,(1+z)$, except in the lowest tomographic bin ($0.1 < z < 0.3$), where it reaches $|\Delta_{\langle z\rangle}| = 0.02\,(1 + z)$. Using the de-biased zPDF method, we find $|\Delta_{\langle z\rangle}| \lesssim 0.01\,(1 + z)$. In that case, the $\sigma_{z_{\rm p}} < 0.3$ selection removes between 20% and 44% of the full KiDS+VIKING-450 sample[4]. If we relax the selection on the photo-z error, as presented in the lower section of Table 1, the bias $\Delta_{\langle z\rangle}$ increases with the de-biased zPDF approach, as found in the simulation. Nonetheless, $\Delta_{\langle z\rangle}$ remains around 1%, which corresponds to an accuracy comparable to that obtain with direct calibration. We note that the zPDF de-biasing technique with the photo-z prior performs significantly better than with the flat prior. Figure 9 illustrates the impact of the photo-z prior in

[3] We limit the risk of incorrect association between the photometric and spectroscopic sources by allowing a maximum angular separation of $0\farcs3$ in the match between the KiDS-VIKING+450 and VVDS/DEEP2 catalogues.

[4] The representation fraction changes in each tomographic bin due to correlations between spec-z and photo-z uncertainties.

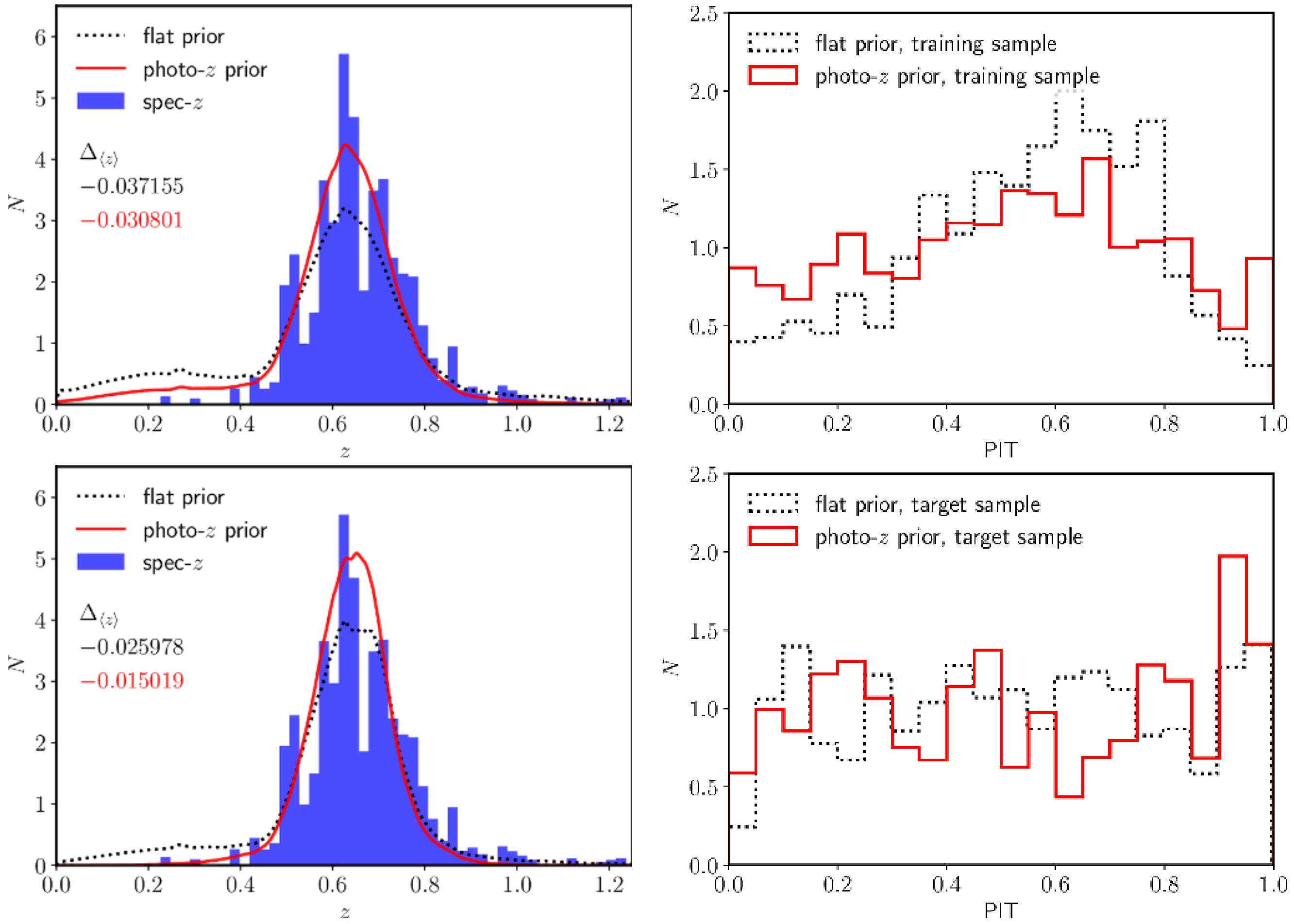

Fig. 9. Same as Fig. 4, except that this refers to real data from the KiDS+VIKING-450 photometric survey and the VVDS-DEEP2 target sample. The sample is selected with a $\sigma_{z_p} < 0.6$ threshold in the photo-z uncertainties.

recovering the shape of the redshift distribution, where we can see a clear improvement below the main mode (bottom-left panel). This result is confirmed in the other tomographic bins.

The depth of the KiDS imaging survey is similar to the one we simulated for DES (5σ sensitivity between 23.6 and 25.1), while the VIKING photometry is much shallower than the *Euclid* one (between 21.2 and 22.7 for VIKING). It is therefore encouraging to find a bias similar to that expected from the simulation in the DES/*Euclid* configuration, even with shallower imaging. We emphasise that our estimate is performed in the worst possible conditions: (1) Our training sample does not cover the same colour and magnitude space as our target sample, as shown in Wright et al. (2020), (2) the photometric calibration could vary from field to field, and (3) some failures in the spec-z target sample could bias the mean redshift considered as the truth. We know that a fraction of the target spec-z could include catastrophic failures, possibly biasing our estimate of $\langle z\rangle_{\rm true}$. Indeed, flag 3 in VVDS and DEEP2 are expected to be 97% and 95% correct, respectively, suggesting that a few percent of failures may be present in those samples, thereby introducing a bias in the true mean redshift, $\langle z\rangle_{\rm true}$, of more than 0.01, according to Fig. 7. The presence of such a fraction of failures remains difficult to verify. A comparison between duplicated observations in DEEP2 shows that the fraction of failures should be at maximum 1.6% (Newman et al. 2013).

Finally, we note that our various selections on σ_{z_p} prevent us from directly comparing the recovered redshift distributions with those published in Wright et al. (2019) and Joudaki et al. (2020). Indeed, our selection on σ_{z_p} preferentially removes the faintest galaxies from the sample, thus shifting the intrinsic redshift distribution towards redshifts that are lower than expected for the full KiDS+VIKING-450 sample.

7. Summary and conclusion

This paper investigates the possibility of measuring the mean redshift $\langle z\rangle$ of a target sample of galaxies, in ten tomographic bins from $z = 0.2$ to $z = 2.2$, with an accuracy of $|\Delta_{\langle z\rangle}| < 0.002\,(1+z)$, as stipulated by the *Euclid* mission requirements on cosmic shear analysis. Naturally, the conclusions presented here are equally applicable to all current and future surveys where redshift calibration is a relevant challenge.

We applied two approaches, which are foreseen for the *Euclid* mission: a direct calibration of $\langle z\rangle$ with a spectroscopic training sample and the combination of individual zPDFs to reconstruct the underlying redshift distribution. This paper analyses in detail several factors that could impact these approaches and provides recommendations on how to apply them successfully.

We used the Horizon-AGN hydrodynamical simulation (Dubois et al. 2014), which allows a large diversity of modelled SEDs, and created 18 mock *Euclid*-like catalogues with different realisations of the photometric noise. We simulated two possible configurations, which should encompass the range of sensitivities of future imaging available for *Euclid*: (1) a shallow configuration combining DES and *Euclid* and (2) a deep configuration combining LSST and *Euclid*. We measured the photo-z of the simulated galaxies using the template-fitting code `LePhare`, as performed in Laigle et al. (2019). Such a procedure produces photometric redshifts with complex zPDFs, realistic biases, and catastrophic failures. We also assumed different characteristics for the spectroscopic training samples associated with the mock catalogues. We considered several selection functions and sample sizes and included possible failures in the spec-z.

We first tested the direct calibration approach, where the redshift distribution is directly estimated from existing spectroscopic redshifts in a training sample, applying necessary weights to match this distribution to the target sample. We find that this approach is efficient in recovering the mean redshift with an accuracy of $0.002\,(1+z)$. The method is successful when based on a representative spectroscopic coverage (uniform or SOM), but the weighting scheme is not sufficient to correct for the heterogeneity in the COSMOS-like training sample at the level required by *Euclid*. This method is stable and robust and does not require deep photometry such as that from LSST. However, we find that the recovered mean redshift is extremely sensitive to the presence of catastrophic failures in spectroscopic redshift measurement. To recover unbiased estimates of $\langle z \rangle$, a careful quality assessment of the spectroscopic redshifts must guarantee a fraction of failures below 0.2%.

We then investigated the possibility of reconstructing the redshift distribution from the zPDF produced by a template-fitting photo-z code. As expected, we find that the quality of the initial zPDF is not sufficient to measure $\langle z \rangle$ with an accuracy better than $|\Delta_{\langle z \rangle}| < 0.01$. We tested the method from Bordoloi et al. (2010) to de-bias the zPDF. We improved it by taking into account an appropriate prior combined with an iterative correction of the zPDF. Our results are summarised below.

- The mean redshift accuracy inferred from the de-biased zPDF is systematically improved when compared to the one inferred from the initial zPDF (by up to a factor ten).
- This method is weakly sensitive to the fraction of spec-z failures.
- Imaging depth is the primary factor in determining the effectiveness of the de-biasing technique. We reach the *Euclid* requirement when combining *Euclid* and LSST ground-based images.
- Insufficient imaging depth can be compensated for by selecting well-peaked zPDFs, but it introduces considerable losses to the target sample number density. A balance should therefore be established between the accuracy of $\langle z \rangle$ and the statistical signal of the cosmic shear analysis.

We tested the two approaches on real datasets from COSMOS and KiDS+VIKING-450 and confirm that a high signal-to-noise in the photometry is essential for an accurate estimate of $\langle z \rangle$ using the de-biased zPDF approach. In the less favourable case (KiDS+VIKING-450), where the photometric sample and a spec-z target sample are approximately of equal depth, we reach an accuracy of around $0.01\,(1+z)$ on $\langle z \rangle$, as expected from the simulation and other works (e.g., Wright et al. 2020). We confirm the trends observed in the simulation and find that including the prior in the de-biasing technique produces significantly better results.

We conclude that both methods could foreseeably provide independent and accurate inferences of tomographic bin mean redshifts for *Euclid*. We find that the current *Euclid* baseline to measure $\langle z \rangle$ with a direct calibration approach and a SOM training sample is robust with respect to the imaging survey depth. However, we recommend that training samples, such as C3R2 (Masters et al. 2019), ensure a purity level above 99.8%. We also find that the sum of the de-biased zPDFs could be sufficient to measure $\langle z \rangle$ at the *Euclid* requirement with ongoing spectroscopic surveys. However, we recommend this method only in areas covered with deep optical data. The two methods should be applied simultaneously with the current planning of the *Euclid* survey to provide complementary and independent estimates of $\langle z \rangle$.

Finally, our work suffers several limitations that we still need to investigate. We have neglected the catastrophic failures within the photo-z sample created by misclassified stars or AGN or by the galaxy blending. A residual contamination of these populations in the tomographic bins could affect both approaches to redshift calibration. Moreover, we have not considered sample variance effects since the Horizon-AGN simulation covers only 1 deg^2. We would benefit from a larger simulated area to test the impact of sample variance. Nonetheless, our results here present a largely positive outlook for the challenge of tomographic redshift calibration within *Euclid*.

Acknowledgements. We thank the OU-PHZ of *Euclid* for all the useful discussions along these years. OI acknowledges the funding of the French Agence Nationale de la Recherche for the project 'SAGACE'. NM acknowledges support from a CNES fellowship. H. Hildebrandt is supported by a Heisenberg grant of the Deutsche Forschungsgemeinschaft (Hi 1495/5-1) as well as an ERC Consolidator Grant (No. 770935). A.H. Wright is supported by the ERC Consolidator Grant (No. 770935). This work relied on the HPC resources of CINES (Jade) under the allocation 2013047012 and c2014047012 made by GENCI and on the Horizon Cluster hosted by Institut d'Astrophysique de Paris. ID acknowledges that he received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 896225. We warmly thank S. Rouberol for running the cluster on which the simulation was post-processed. This research is also partly supported by the Centre National d'Etudes Spatiales (CNES). We would also like to recognise the contributions from all of the members of the COSMOS team who helped in obtaining and reducing the large amount of multi-wavelength and spectroscopic data. Based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 177.A-3016, 177.A-3017, 177.A-3018, 179.A-2004, and on data products produced by the KiDS consortium. The KiDS production team acknowledges support from: Deutsche Forschungsgemeinschaft, ERC, NOVA and NWO-M grants; Target; the University of Padova, and the University Federico II (Naples). SA thank the support PRIN MIUR2015 "Cosmology and Fundamental Physics: Illuminating the Dark Universe with Euclid". The *Euclid* Consortium acknowledges the European Space Agency and a number of agencies and institutes that have supported the development of Euclid, in particular the Academy of Finland, the Agenzia Spaziale Italiana, the Belgian Science Policy, the Canadian *Euclid* Consortium, the Centre National d'Etudes Spatiales, the Deutsches Zentrum für Luft- und Raumfahrt, the Danish Space Research Institute, the Fundação para a Ciencia e a Tecnologia, the Ministerio de Economia y Competitividad, the National Aeronautics and Space Administration, the Netherlandse Onderzoekschool Voor Astronomie, the Norwegian Space Agency, the Romanian Space Agency, the State Secretariat for Education, Research and Innovation (SERI) at the Swiss Space Office (SSO), and the United Kingdom Space Agency. A complete and detailed list is available on the *Euclid* website (`http://www.euclid-ec.org`).

References

Abbott, T. M. C., Abdalla, F. B., Allam, S., et al. 2018, ApJS, 239, 18
Abdalla, F. B., Amara, A., Capak, P., et al. 2008, MNRAS, 387, 969
Arnouts, S., Moscardini, L., Vanzella, E., et al. 2002, MNRAS, 329, 355
Aubert, D., Pichon, C., & Colombi, S. 2004, MNRAS, 352, 376
Benítez, N. 2000, ApJ, 536, 571
Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476
Bordoloi, R., Lilly, S. J., & Amara, A. 2010, MNRAS, 406, 881
Brodwin, M., Lilly, S. J., Porciani, C., et al. 2006, ApJS, 162, 20
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000

Caldwell, R. R., Dave, R., & Steinhardt, P. J. 1998, Phys. Rev. Lett., 80, 1582
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Chabrier, G. 2003, PASP, 115, 763
Comparat, J., Richard, J., Kneib, J.-P., et al. 2015, A&A, 575, A40
Cunha, C. E., Lima, M., Oyaizu, H., Frieman, J., & Lin, H. 2009, MNRAS, 396, 2379
Daddi, E., Cimatti, A., Renzini, A., et al. 2004, ApJ, 617, 746
Davidzon, I., Laigle, C., Capak, P. L., et al. 2019, MNRAS, 489, 4817
Dawid, A. 1984, J. R. Stat. Soc., 147, 278
De Santis, C., Grazian, A., Fontana, A., & Santini, P. 2007, New Astron., 12, 271
Dubois, Y., Devriendt, J., Slyz, A., & Teyssier, R. 2012, MNRAS, 420, 2662
Dubois, Y., Pichon, C., Welker, C., et al. 2014, MNRAS, 444, 1453
Fosalba, P., Crocce, M., Gaztañaga, E., & Castander, F. J. 2015, MNRAS, 448, 2987
Fotopoulou, S., & Paltani, S. 2018, A&A, 619, A14
Graham, M. L., Connolly, A. J., Wang, W., et al. 2020, AJ, 159, 258
Guglielmo, V., Saglia, R., Castander, F. J., et al. 2020, A&A, 642, A192
Haardt, F., & Madau, P. 1996, ApJ, 461, 20
Hartley, W. G., Chang, C., Samani, S., et al. 2020, MNRAS, 496, 4769
Hayward, C. C., Sparre, M., Chapman, S. C., et al. 2021, MNRAS, 502, 2922
Hikage, C., Oguri, M., Hamana, T., et al. 2019, PASJ, 71, 43
Hildebrandt, H., Erben, T., Kuijken, K., et al. 2012, MNRAS, 421, 2355
Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, MNRAS, 465, 1454
Hildebrandt, H., Köhlinger, F., van den Busch, J. L., et al. 2020, A&A, 633, A69
Hu, W. 1999, ApJ, 522, L21
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841
Ilbert, O., Capak, P., Salvato, M., et al. 2009, ApJ, 690, 1236
Ilbert, O., McCracken, H. J., Le Fèvre, O., et al. 2013, A&A, 556, A55
Joudaki, S., Hildebrandt, H., Traykova, D., et al. 2020, A&A, 638, L1
Kacprzak, T., Herbel, J., Nicola, A., et al. 2020, Phys. Rev. D, 101, 082003
Kashino, D., Silverman, J. D., Sanders, D., et al. 2019, ApJS, 241, 10
Kaviraj, S., Laigle, C., Kimm, T., et al. 2017, MNRAS, 467, 4739
Kilbinger, M. 2015, Rep. Progr. Phys., 78, 086901
Kilbinger, M., Fu, L., Heymans, C., et al. 2013, MNRAS, 430, 2200
Kohonen, T. 1982, Biol. Cybern., 43, 59
Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, ApJS, 192, 18
Kuijken, K., Heymans, C., Dvornik, A., et al. 2019, A&A, 625, A2
Laidler, V. G., Papovich, C., Grogin, N. A., et al. 2007, PASP, 119, 1325
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Laigle, C., Davidzon, I., Ilbert, O., et al. 2019, MNRAS, 486, 5104
Lang, D., Hogg, D. W., & Mykytyn, D. 2016, Astrophys. Source Code Libr., [record ascl:1604.008]
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, A&A, 439, 845
Le Fèvre, O., Tasca, L. A. M., Cassata, P., et al. 2015, A&A, 576, A79
Lilly, S. J., Le Fèvre, O., Renzini, A., et al. 2007, ApJS, 172, 70
Lima, M., Cunha, C. E., Oyaizu, H., et al. 2008, MNRAS, 390, 118
LSST Science Collaboration (Abell, P. A., et al.) 2009, ArXiv e-prints [arXiv:0912.0201]
Ma, Z., Hu, W., & Huterer, D. 2006, ApJ, 636, 21
Mandelbaum, R. 2018, ARA&A, 56, 393
Masters, D., Capak, P., Stern, D., et al. 2015, ApJ, 813, 53
Masters, D. C., Stern, D. K., Cohen, J. G., et al. 2017, ApJ, 841, 111
Masters, D. C., Stern, D. K., Cohen, J. G., et al. 2019, ApJ, 877, 81
Ménard, B., Scranton, R., Schmidt, S., et al. 2013, ArXiv e-prints [arXiv:1303.4722]
Merlin, E., Bourne, N., Castellano, M., et al. 2016, A&A, 595, A97
Merloni, A., Predehl, P., Becker, W., et al. 2012, ArXiv e-prints [arXiv:1209.3114]
Newman, J. A. 2008, ApJ, 684, 88
Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5
Perlmutter, S., Aldering, G., Goldhaber, G., et al. 1999, ApJ, 517, 565
Pichon, C., Thiébaut, E., Prunet, S., et al. 2010, MNRAS, 401, 705
Polletta, M., Tajer, M., Maraschi, L., et al. 2007, ApJ, 663, 81
Prevot, M. L., Lequeux, J., Prevot, L., Maurice, E., & Rocca-Volmerange, B. 1984, A&A, 132, 389
Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, AJ, 116, 1009
Salvato, M., Ilbert, O., Hasinger, G., et al. 2011, ApJ, 742, 61
Salvato, M., Ilbert, O., & Hoyle, B. 2019, Nat. Astron., 3, 212
Schaerer, D., & de Barros, S. 2009, A&A, 502, 423
Sheth, R. K. 2007, MNRAS, 378, 709
Sheth, R. K., & Rossi, G. 2010, MNRAS, 403, 2137
Spergel, D., Gehrels, N., Baltay, C., et al. 2015, ArXiv e-prints [arXiv:1503.03757]
Straatman, C. M. S., van der Wel, A., Bezanson, R., et al. 2019, VizieR Online Data Catalog: J/ApJS/239/27
Sutherland, R. S., & Dopita, M. A. 1993, ApJS, 88, 253
Troxel, M. A., MacCrann, N., Zuntz, J., et al. 2018, Phys. Rev. D, 98, 043528
Urrutia, T., Wisotzki, L., Kerutt, J., et al. 2019, A&A, 624, A141
van der Wel, A., Noeske, K., Bezanson, R., et al. 2016, ApJS, 223, 29
Wright, A. H., Hildebrandt, H., Kuijken, K., et al. 2019, A&A, 632, A34
Wright, A. H., Hildebrandt, H., van den Busch, J. L., & Heymans, C. 2020, A&A, 637, A100

[1] Aix-Marseille Univ., CNRS, CNES, LAM, Marseille, France
e-mail: olivier.ilbert@lam.fr
[2] Ruhr-Universität Bochum, Astronomisches Institut, German Centre for Cosmological Lensing, Universitätsstr. 150, 44801 Bochum, Germany
[3] Department of Astronomy, University of Geneva, Ch. d'Écogia 16, 1290 Versoix, Switzerland
[4] Institut d'Astrophysique de Paris, 98bis boulevard Arago, 75014 Paris, France
[5] Cosmic Dawn Center (DAWN), Niels Bohr Institute, University of Copenhagen, Vibenshuset, Lyngbyvej 2, 2100 Copenhagen, Denmark
[6] Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA
[7] Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK
[8] Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK
[9] INAF-Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, Italy
[10] INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy
[11] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK
[12] IFPU, Institute for Fundamental Physics of the Universe, Via Beirut 2, 34151 Trieste, Italy
[13] SISSA, International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, TS, Italy
[14] INFN, Sezione di Trieste, Via Valerio 2, 34127 Trieste, TS, Italy
[15] INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, 34131 Trieste, Italy
[16] Universidad de la Laguna, 38206 San Cristóbal de La Laguna, Tenerife, Spain
[17] Instituto de Astrofísica de Canarias, Calle Vía Làctea s/n, 38204 San Cristóbal de la Laguna, Tenerife, Spain
[18] Dipartimento di Fisica e Astronomia, Universitá di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy
[19] INFN-Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[20] INAF-Osservatorio Astronomico di Padova, Via dell'Osservatorio 5, 35122 Padova, Italy
[21] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstrasse 1, 81679 München, Germany
[22] Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, 85748 Garching, Germany
[23] INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese, TO, Italy
[24] Dipartimento di Fisica – Sezione di Astronomia, Universitá di Trieste, Via Tiepolo 11, 34131 Trieste, Italy
[25] Université de Paris, CNRS, Astroparticule et Cosmologie, 75006 Paris, France
[26] INFN-Sezione di Roma Tre, Via della Vasca Navale 84, 00146 Roma, Italy
[27] Department of Mathematics and Physics, Roma Tre University, Via della Vasca Navale 84, 00146 Rome, Italy
[28] INAF-Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monteporzio Catone, Italy
[29] INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy

[30] INFN-Bologna, Via Irnerio 46, 40126 Bologna, Italy
[31] Dipartimento di Fisica e Scienze della Terra, Universitá degli Studi di Ferrara, Via Giuseppe Saragat 1, 44122 Ferrara, Italy
[32] INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, 40129 Bologna, Italy
[33] Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, 31400 Toulouse, France
[34] INFN-Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy
[35] Dipartimento di Fisica, Universitá degli Studi di Torino, Via P. Giuria 1, 10125 Torino, Italy
[36] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France
[37] INAF-IASF Milano, Via Alfonso Corti 12, 20133 Milano, Italy
[38] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra, Barcelona, Spain
[39] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda, 1349-018 Lisboa, Portugal
[40] AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, 91191 Gif-sur-Yvette, France
[41] Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain
[42] Institut d'Estudis Espacials de Catalunya (IEEC), Carrer Gran Capitá 2-4, 08034 Barcelona, Spain
[43] Observatoire de Sauverny, Ecole Polytechnique Fédérale de Lausanne, 1290 Versoix, Switzerland
[44] Department of Physics "E. Pancini", University Federico II, Via Cinthia 6, 80126 Napoli, Italy
[45] INFN Section of Naples, Via Cinthia 6, 80126 Napoli, Italy
[46] INAF-Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125 Firenze, Italy
[47] Centre National d'Etudes Spatiales, Toulouse, France
[48] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[49] European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati, Roma, Italy
[50] ESAC/ESA, Camino Bajo del Castillo s/n, Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain
[51] Univ. Lyon, Univ. Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, 69622 Villeurbanne, France
[52] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, 1749-016 Lisboa, Portugal
[53] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[54] Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK
[55] INFN-Padova, Via Marzolo 8, 35131 Padova, Italy
[56] University of Lyon, UCB Lyon 1, CNRS/IN2P3, IUF, IP2I, Lyon, France
[57] School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK
[58] Aix-Marseille Univ., CNRS/IN2P3, CPPM, Marseille, France
[59] Department of Physics, University of Helsinki, PO Box 64, 00014 Helsinki, Finland
[60] Dipartimento di Fisica "Aldo Pontremoli", Universitá degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy
[61] INFN-Sezione di Milano, Via Celoria 16, 20133 Milano, Italy
[62] Institute of Theoretical Astrophysics, University of Oslo, PO Box 1029, Blindern 0315, Oslo, Norway
[63] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[64] von Hoerner & Sulger GmbH, SchloßPlatz 8, 68723 Schwetzingen, Germany
[65] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany
[66] Department of Physics and Helsinki Institute of Physics, University of Helsinki, Gustaf Hällströmin Katu 2, 00014 Helsinki, Finland
[67] Université de Genève, Département de Physique Théorique and Centre for Astroparticle Physics, 24 Quai Ernest-Ansermet, 1211 Genève 4, Switzerland
[68] NOVA Optical Infrared Instrumentation Group at ASTRON, Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands
[69] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
[70] Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK
[71] Institut für Theoretische Physik, University of Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany
[72] Zentrum für Astronomie, Universität Heidelberg, Philosophenweg 12, 69120 Heidelberg, Germany
[73] INAF-IASF Bologna, Via Piero Gobetti 101, 40129 Bologna, Italy
[74] Université de Paris, 75013 Paris, France
[75] LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Université, 75014 Paris, France
[76] CEA Saclay, DFR/IRFU, Service d'Astrophysique, Bât. 709, 91191 Gif-sur-Yvette, France
[77] IRFU, CEA, Université Paris-Saclay, 91191 Gif-sur-Yvette Cedex, France
[78] ICC&CEA, Department of Physics, Durham University, South Road, DH1 3LE Durham, UK
[79] Department of Physics and Astronomy, University of Aarhus, Ny Munkegade 120, 8000 Aarhus C, Denmark
[80] Space Science Data Center, Italian Space Agency, Via del Politecnico snc, 00133 Roma, Italy
[81] Institute of Space Science, Bucharest 077125, Romania
[82] Institute for Computational Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[83] Dipartimento di Fisica e Astronomia "G. Galilei", Universitá di Padova, Via Marzolo 8, 35131 Padova, Italy
[84] Departamento de Física, FCFM, Universidad de Chile, Blanco Encalada 2008, Santiago, Chile
[85] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain
[86] Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, 30202 Cartagena, Spain
[87] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands
[88] Department of Physics, University of Jyväskylä, PO Box 35 (YFL), 40014 Jyväskylä, Finland
[89] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

Appendix A: Idealised test of the de-biasing procedure

In this appendix, we present how we generated a simplified mock catalogue in comparison to the one presented in Sect. 2. We still used the mock Horizon-AGN catalogue. Rather than using the photo-z produced by LePhare, however, we generated an idealised photo-z. We randomised the true redshift assuming a Gaussian distribution with $\sigma = \sigma_{\rm true}$, where $\sigma_{\rm true}$ is defined as the median value of the LePhare photo-z errors. We then biased these photo-z by applying a systematic shift of $\Delta_{z_{\rm p}} = -0.05$. We associated a likelihood with each galaxy defined as:

$$\mathcal{L}(z) = \frac{1}{A\,\sigma_{\rm true}\,\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - z_{\rm p}}{A\,\sigma_{\rm true}}\right)^2\right], \tag{A.1}$$

where the factor A allows us to mimic an underestimation (overestimation) of the photo-z uncertainties if $A < 1$ ($A > 1$). In this way, we can check, using a simplified simulation, if we are able to recover the true mean redshift despite having a bias in the photo-z and their associated likelihood.

We applied the same method as described in Sect. 4.3 to recover the mean redshift, assuming a flat prior. We selected galaxies in a tomographic bin at $0.6 < z_{\rm p} < 0.8$. Two examples are given in Fig. A.1. The top (bottom) panels assume $A = 0.7$ ($A = 1.5$), that is to say, that photo-z errors are underestimated (overestimated).

We find that as long as $A > 1$, the method is efficient in recovering the mean redshift. However, if the original zPDFs are too narrow ($A < 1$), the final correction is unstable. We find the same result by testing several values of A and several values of the bias. Therefore, we conclude that photo-z errors should be preferentially overestimated in the application of the de-biased zPDF method.

As a result, when applying our template-fitting code to the simulated Horizon-AGN galaxies, we simply multiply the flux uncertainties by a constant factor to ensure that we are working in this regime. Specifically, for comparison to the photo-z measured by Laigle et al. (2019), we multiply the flux uncertainties by a factor of 1.5 and impose a minimal error of $\Delta m = 0.01$ in each band.

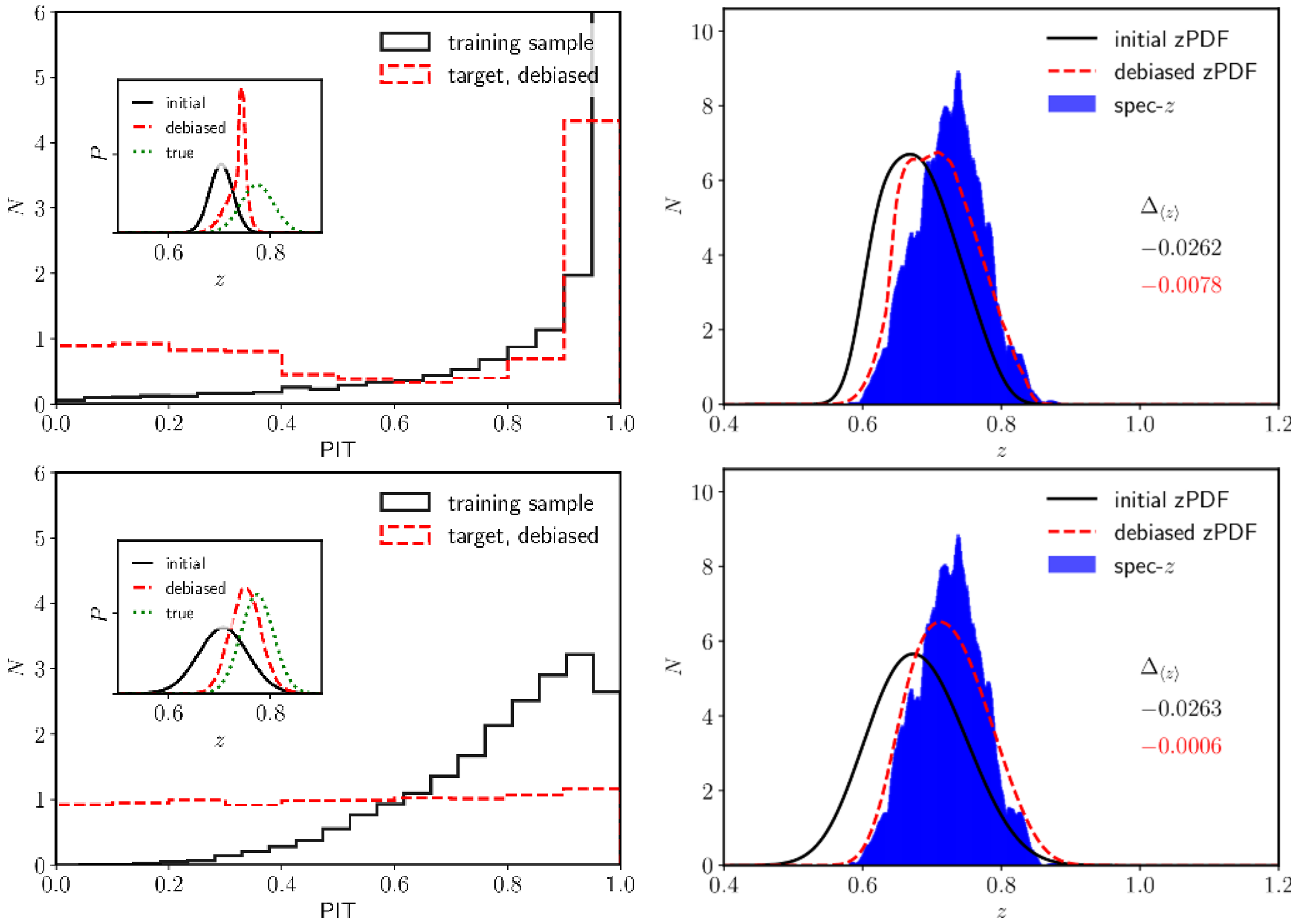

Fig. A.1. Example of PIT distribution (*left*) and redshift distribution (*right*) for a tomographic bin selected at $0.6 < z_{\rm p} < 0.8$. *Top* and *bottom panels*: assume photo-z errors that are underestimated ($A = 0.7$) and overestimated ($A = 1.5$), respectively. The PIT distribution used to correct the zPDF is shown with the solid black line. The inset shows an example of the de-biased zPDF for one galaxy (selected randomly). The resulting PIT distribution, after de-biasing, is shown in dashed red. The true $N(z)$ is shown with the blue histogram in the *right panels*. The $N(z)$ reconstructed using the initial and the de-biased zPDFs are shown with black solid lines and red dashed lines, respectively.

Appendix B: Mean redshift of the LEGA-C survey in COSMOS

The goal in this section is to retrieve the mean redshift of the LEGA-C galaxies (van der Wel et al. 2016) selected in the tomographic bin $0.7 < z_\mathrm{p} < 0.9$. We based our estimate of $\langle z \rangle$ on the COSMOS photometry and associated spec-z (excluding LEGA-C spec-z from the training). Then, we compared the estimated mean redshift with the true one (known from LEGA-C spec-z). In such a configuration, the photometry is much deeper than the selection limit of the target sample.

The COSMOS photometry. We used the photometric catalogue from Laigle et al. (2016), but keeping only the ten broad bands: u, B, V, r, i, z, Y, J, H, and K. We adopted the exact same method as the one described in Sect. 2.2 to compute the photo-z. As described in Sect. 4.3, we inflated our photometric flux uncertainties within the input photometric catalogue by a factor of two to allow for better de-biasing.

LEGA-C target sample. We selected a spectroscopic sample that was as robust as possible to ensure that the uncertainty on the mean redshift of the target sample (considered as the truth) is known with an accuracy better than 0.002. The LEGA-C spectroscopic survey in the COSMOS field provides such a target sample. This spectroscopic sample is built using the high-resolution ($R = 3000$) mode of the VIMOS spectrograph, targeting galaxies at $0.6 < z < 1$ selected in the K_s-band to have a stellar mass $M_\star > 10^{10}\,M_\odot$. Given the resolution and the S/N reached by the LEGA-C spectra (with 20 h of exposure per spectrum) and the numerous lines detected, we can safely assume that this sample does not include any catastrophic spectroscopic failures. We matched the LEGA-C Data Release 2 galaxies (Straatman et al. 2019) to the COSMOS2015 catalogue on-sky, allowing a maximum angular separation of $0\farcs2$ in the association. This reduced the risk of incorrectly associating spectra with our COSMOS2015 photometry. Our LEGA-C target sample thus contains 1213 galaxies, with a median i-band magnitude of 21.45.

The COSMOS training sample. Since the constraint in terms of completeness and purity is less stringent for the training sample, we randomly chose 50% of all the spec-z available in COSMOS, irrespective of magnitude. We removed all the LEGA-C sources from the training sample and combined the spec-z from multiple surveys, namely: zCOSMOS-Bright and Faint (Lilly et al. 2007), Fiber-Multi Object Spectrograph (FMOS; Kashino et al. 2019), and C3R2 (Masters et al. 2019). We selected only spectra with either ‘high confidence’ or ‘certain’ redshift confidence flags (corresponding to flags 3–4 in the VVDS redshift confidence flagging system in Le Fèvre et al. 2005) in order to select only the most reliable redshifts for our training set. Still, the magnitude and colour distributions differed between the training and the target samples. We thus applied a weight to each galaxy of the training sample to reproduce the global properties of the target sample. Those weights were derived by projecting the target sample over the SOM, as described in Sect. 3 for the COSMOS-like sample. We constructed our SOM here using the magnitudes, colours, and photo-z associated with the training sample sources. We adopted a 10×10 SOM, smaller than the one used in Horizon-AGN, because of the limited size of the target sample.

Application. We selected all sources with photo-z in the range $0.7 < z_\mathrm{p} < 0.9$ (we chose this redshift range since it needs to overlap with LEGA-C). We created 300 realisations with a random selection of the training sources. The target sample consisted of 493 galaxies, of which around 5% have $\sigma_{z_\mathrm{p}} > 0.3$ and were subsequently removed. We estimated the mean redshift of the target sample using the direct calibration, direct zPDF, and de-biased zPDF approaches, and compared these with the true $\langle z \rangle$ of the target sample. For the direct calibration approach, we obtain a bias of $\mu_{\Delta z} = 0.00032$ and a scatter of $\sigma_{\Delta z} = 0.00135$, an accuracy well within the *Euclid* requirement. Secondly, we estimated $\langle z \rangle$ using the initial zPDF without de-biasing. We obtain a mean redshift biased by $\mu_{\Delta z} > -0.013$, which is six times larger than the *Euclid* requirement. Finally, we de-bias the zPDF using the PIT distribution as discussed in Sect. 4.3. In that case, we obtain a mean redshift with a bias of $\mu_{\Delta z} = -0.00046$ ($\mu_{\Delta z} = -0.00008$) and a scatter of $\sigma_{\Delta z} = 0.00073$ ($\sigma_{\Delta z} = 0.00074$) assuming the photo-$z$ (flat) prior. Therefore, in the case of a target sample associated with much deeper photometry, we reach the $0.002\,(1 + z)$ accuracy requirement of *Euclid*, using either the direct calibration or de-biased zPDF approaches.