Euclid preparation. X. The Euclid photometric-redshift challenge

G. Desprez, S. Paltani, J. Coupon, I. Almosallam, A. Alvarez-Ayllon, V. Amaro, M. Brescia, M. Brodwin, S. Cavuoti, J. de Vicente-Albendea, et al.

A&A 644, A31 (2020)
https://doi.org/10.1051/0004-6361/202039403

Astronomy & Astrophysics

Euclid preparation

X. The *Euclid* photometric-redshift challenge

Euclid Collaboration: G. Desprez[1,⋆], S. Paltani[1], J. Coupon[1], I. Almosallam[2,3,4], A. Alvarez-Ayllon[1], V. Amaro[5], M. Brescia[6], M. Brodwin[7], S. Cavuoti[6,8,9], J. De Vicente-Albendea[10], S. Fotopoulou[11], P. W. Hatfield[12], W. G. Hartley[1], O. Ilbert[13], M. J. Jarvis[12], G. Longo[8,9], M. M. Rau[14], R. Saha[7], J. S. Speagle[15], A. Tramacere[1], M. Castellano[16], F. Dubath[1], A. Galametz[1], M. Kuemmel[17], C. Laigle[18], E. Merlin[16], J. J. Mohr[17,19], S. Pilo[16,†], M. Salvato[19], S. Andreon[20], N. Auricchio[21], C. Baccigalupi[22,23,24], A. Balaguera-Antolínez[25,26], M. Baldi[21,27,28], S. Bardelli[21], R. Bender[17,19], A. Biviano[24,29], C. Bodendorf[19], D. Bonino[30], E. Bozzo[1], E. Branchini[16,31,32], J. Brinchmann[33], C. Burigana[28,34,35], R. Cabanac[36], S. Camera[30,37,38], V. Capobianco[30], A. Cappi[21,39], C. Carbone[40], J. Carretero[41], C. S. Carvalho[42], R. Casas[43,44], S. Casas[45], F. J. Castander[43,44], G. Castignani[46], A. Cimatti[27,47], R. Cledassou[48], C. Colodro-Conde[26], G. Congedo[49], C. J. Conselice[50], L. Conversi[51,52], Y. Copin[53], L. Corcione[30], H. M. Courtois[54], J.-G. Cuby[13], A. Da Silva[55,56], S. de la Torre[13], H. Degaudenzi[1], D. Di Ferdinando[28], M. Douspis[57], C. A. J. Duncan[12], X. Dupac[52], A. Ealet[53], G. Fabbian[58], M. Fabricius[19], S. Farrens[45], P. G. Ferreira[59], F. Finelli[21,60], P. Fosalba[43,44], N. Fourmanoit[61], M. Frailis[24], E. Franceschi[21], M. Fumana[40], S. Galeotta[24], B. Garilli[40], W. Gillard[62], B. Gillis[49], C. Giocoli[21,27,28], G. Gozaliasl[63,64], J. Graciá-Carpio[19], F. Grupp[19], L. Guzzo[20,65,66], M. Hailey[67], S. V. H. Haugan[68], W. Holmes[69], F. Hormuth[70], A. Humphrey[33], K. Jahnke[71], E. Keihanen[64], S. Kermiche[62], M. Kilbinger[45,72], C. C. Kirkpatrick[64], T. D. Kitching[67], R. Kohley[52], B. Kubik[54], M. Kunz[73], H. Kurki-Suonio[64], S. Ligori[30], P. B. Lilje[68], I. Lloro[74], D. Maino[40,65,66], E. Maiorano[21], O. Marggraf[75], K. Markovic[69], N. Martinet[13], F. Marulli[21,27,28], R. Massey[76], M. Maturi[77,78], N. Mauri[27,28], S. Maurogordato[79], E. Medinaceli[80], S. Mei[81,82], M. Meneghetti[21,28], R. Benton Metcalf[27,80], G. Meylan[46], M. Moresco[21,27], L. Moscardini[21,27,60], E. Munari[24], S. Niemi[67], C. Padilla[41], F. Pasian[24], L. Patrizii[28], V. Pettorino[45], S. Pires[45], G. Polenta[83], M. Poncet[48], L. Popa[84], D. Potter[85], L. Pozzetti[21], F. Raison[19], A. Renzi[86,87], J. Rhodes[69], G. Riccio[6], E. Rossetti[27], R. Saglia[17,19], D. Sapone[88], P. Schneider[75], V. Scottez[72], A. Secroun[62], S. Serrano[43,44], C. Sirignano[86,87], G. Sirri[28], L. Stanco[86], D. Stern[69], F. Sureau[45], P. Tallada Crespí[10], D. Tavagnacco[24], A. N. Taylor[49], M. Tenti[60], I. Tereno[42,55], R. Toledo-Moreo[89], F. Torradeflot[10], L. Valenziano[21,28], J. Valiviita[64], T. Vassallo[17], M. Viel[22,23,24,29], Y. Wang[90], N. Welikala[49], L. Whittaker[91,92], A. Zacchei[24], G. Zamorani[21], J. Zoubian[62], and E. Zucca[21]

(Affiliations can be found after the references)

Received 11 September 2020 / Accepted 20 October 2020

ABSTRACT

Forthcoming large photometric surveys for cosmology require precise and accurate photometric redshift (photo-z) measurements for the success of their main science objectives. However, to date, no method has been able to produce photo-zs at the required accuracy using only the broad-band photometry that those surveys will provide. An assessment of the strengths and weaknesses of current methods is a crucial step in the eventual development of an approach to meet this challenge. We report on the performance of 13 photometric redshift code single value redshift estimates and redshift probability distributions (PDZs) on a common set of data, focusing particularly on the 0.2−2.6 redshift range that the *Euclid* mission will probe. We designed a challenge using emulated *Euclid* data drawn from three photometric surveys of the COSMOS field. The data was divided into two samples: one calibration sample for which photometry and redshifts were provided to the participants; and the validation sample, containing only the photometry to ensure a blinded test of the methods. Participants were invited to provide a redshift single value estimate and a PDZ for each source in the validation sample, along with a rejection flag that indicates the sources they consider unfit for use in cosmological analyses. The performance of each method was assessed through a set of informative metrics, using cross-matched spectroscopic and highly-accurate photometric redshifts as the ground truth. We show that the rejection criteria set by participants are efficient in removing strong outliers, that is to say sources for which the photo-z deviates by more than $0.15(1 + z)$ from the spectroscopic-redshift (spec-z). We also show that, while all methods are able to provide reliable single value estimates, several machine-learning methods do not manage to produce useful PDZs. We find that no machine-learning method provides good results in the regions of galaxy color-space that are sparsely populated by spectroscopic-redshifts, for example $z > 1$. However they generally perform better than template-fitting methods at low redshift ($z < 0.7$), indicating that template-fitting methods do not use all of the information contained in the photometry. We introduce metrics that quantify both photo-z precision and completeness of the samples (post-rejection), since both contribute to the final figure of merit of the science goals of the survey (e.g., cosmic shear from *Euclid*). Template-fitting methods provide the best results in these metrics, but we show that a combination of template-fitting results and machine-learning results with rejection criteria can outperform any individual method. On this basis, we argue that further work in identifying how to best select between machine-learning and template-fitting approaches for each individual galaxy should be pursued as a priority.

Key words. galaxies: distances and redshifts – surveys – techniques: miscellaneous – catalogs

⋆ Corresponding author: G. Desprez, e-mail: Guillaume.Desprez@unige.ch
† Deceased.

1. Introduction

The estimation of galaxy redshifts through their photometry, or photometric redshifts (photo-zs), has evolved significantly since the concept was first proposed. The earliest attempts to determine redshifts from photometry used empirical relations (e.g., Baum 1962; Loh & Spillar 1986; Connolly et al. 1995), which then evolved into template-fitting of the photometry (e.g., Puschell et al. 1982; Koo 1985; Lanzetta et al. 1996; Arnouts et al. 1999; Bolzonella et al. 2000). More recently, machine-learning algorithms have been used, based purely on photometry (e.g., Firth et al. 2003; Tagliaferri et al. 2003; Collister & Lahav 2004), possibly combining photometric and morphological information (e.g., Way et al. 2009; Singal et al. 2011; Gomes et al. 2018; Soo et al. 2018), and even directly fed with image cutouts of the sources (e.g., D'Isanto & Polsterer 2018; Pasquet et al. 2019). Photometric redshifts were first used to complement spectroscopic-redshifts (spec-z) when the latter were not available, and they subsequently became a major tool used in modern cosmological surveys to compute redshifts for large numbers of sources. For instance, the Dark Energy Survey (DES; Flaugher 2005), the Kilo-Degree Survey (KiDS; de Jong et al. 2013), the Hyper Suprime Cam Subaru Strategic Program (HSC-SSP; Aihara et al. 2018), the *Euclid* survey (Laureijs et al. 2011), the *Vera C. Rubin* Observatory Legacy Survey of Space and Time (LSST; Ivezić et al. 2019), and the Roman Space Telescope survey (Akeson et al. 2019) all rely or will rely on photo-zs to carry out their main science goals. Salvato et al. (2019) present a review of the various ways to compute photo-zs and the challenges these large surveys face.

For cosmological applications, the quality of photo-z measurements is important, since constraints on cosmological parameters obtained by photometric surveys depend on their precision and their accuracy. The performance of photo-z determination depends on several factors (e.g., the set of filters and their depths, the quality of the photometry, the correction of observational effects, etc.), among which the algorithm plays a key role. For this reason, a large variety of photo-z codes have been developed using different approaches for the problem, and a great deal of work is ongoing to improve these methods.

Tests comparing the results of several methods can be carried out to assess state-of-the-art algorithm performance and to identify possible improvements. Such tests have been performed on different sets of data, including: Hogg et al. (1998) on the *Hubble* Deep Field North; Hildebrandt et al. (2010) on the photo-z Accuracy Testing (PHAT) contest based on simulations and Great Observatories Origins Deep Survey (GOODS; Giavalisco et al. 2004) data; Abdalla et al. (2011) on the SDSS-DR6 Luminous Red Galaxies sample; Dahlen et al. (2013) on the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS; Grogin et al. 2011); Tanaka et al. (2018) on the HSC-SSP data-release 1; and Schmidt et al. (2020) on simulated data.

The *Euclid* survey (Laureijs et al. 2011) is a large photometric and spectroscopic survey, which is planned to cover 15 000 deg^2 of the northern and southern extragalactic sky with a 1.2 m-diameter space telescope. *Euclid*'s main goal is to investigate the Universe's accelerating expansion through two main probes, baryonic acoustic oscillations and weak-lensing tomography. The latter probe requires determination of the shapes and redshifts of galaxies. The measurement of source shapes will be performed using a wide visible band (VIS) covering 540−920 nm. For the determination of the photo-zs, *Euclid* will also perform near-infrared (NIR) photometric observations, in Y, J, and H bands (960−2000 nm), complemented by optical ground-based external observations (EXT) in u, g, r, i and z bands. Laureijs et al. (2011) present the requirements that the *Euclid* photo-zs must meet in order to achieve the desired figure of merit (FoM) for the science goals. The choice of the methods to derive the photo-zs is driven by these requirements. Therefore, the *Euclid* photo-z team has designed a test for photo-z methods using a photometry and filter set defined specifically for *Euclid*. Several photo-z codes, most of them being developed by members of the *Euclid* Collaboration, have been applied to a realistic set of *Euclid*-like data obtained from images of the COSMOS field (Scoville et al. 2007). This field was chosen because of its large collection of spectroscopic-redshifts, required by machine-learning algorithms to perform efficiently.

As in Hogg et al. (1998) and Hildebrandt et al. (2010), we have set up a blind test of the performance of the photo-z methods. They are evaluated using standard estimators, as well as new estimators defined specifically for *Euclid*. However, because the final complementary optical-photometry data sets are expected to be deeper and cover a broader wavelength range in the late stages of the *Euclid* mission than those available here, we do not expect to meet the photo-z requirements; we can only compare the relative performance of the different algorithms. In this challenge, we focus on the precision (the scatter) of the results and the fraction of catastrophic failures, but not on the accuracy (the bias) of the photo-zs. We assume that the *Euclid* photo-zs can be calibrated and therefore that the bias can be removed, for instance using the complete calibration of the color-redshift relation (C3R2; Masters et al. 2015, 2017, 2019; Euclid Collaboration 2020). The precision of photo-zs is nevertheless extremely important for the success of tomographic analyses, because the scatter makes the bins overlap in true-redshift space; hence the larger the scatter, the larger the degeneracy between the weak-lensing signal in the different bins. If too large, this degeneracy would effectively prevent us from studying the evolution of the dark-energy properties across the different epochs of the Universe.

The point of this challenge is first to help the *Euclid* photo-z team to define the strategy of the *Euclid* photo-z pipeline to achieve the photo-z requirements. It also aims to provide clues on ways to improve photo-z method performance by comparing the pros and cons of different approaches.

2. Data

We built a *Euclid*-like wide-survey data set from real photometric data matched as far as possible to the characteristics of the future *Euclid* survey. However, some unavoidable differences are present. First the broad VIS band does not exist in any other survey and thus cannot be simulated from existing data. Also, the ground-based optical data we have used (the DES survey; see Sect. 2.1.2) do not contain any u-band observations, although we expect to have such observations over most of the *Euclid* survey. In addition, deeper ground-based data are expected to be available in the late stage of the survey. Finally, the available NIR images have a lower resolution than *Euclid* will have. For these reasons, this challenge cannot be interpreted as a test of the absolute performance of the photo-z codes, but only as a comparison of the different algorithm under similar conditions.

2.1. Images

The data set is composed of mosaics in eight different bands (g, r, i, z, Y, J, H, and VIS-like) from three different surveys

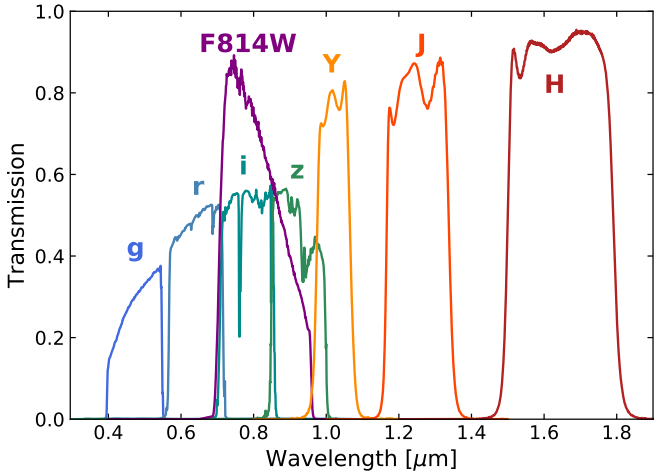

Fig. 1. Transmission curves of the eight filters used in the challenge. The effects of instrumental throughput and atmosphere are included in the transmission.

of the COSMOS field. The area covered by the images is $\sim 1.2 \times 1.2\,\mathrm{deg}^2$. The transmission curves of the filters in these bands are shown in Fig. 1. All the mosaics have been rescaled to the same pixel scale (i.e., 0.″1 per pixel). Table 1 shows the properties of the different mosaics.

2.1.1. VIS-like image

The VIS-like mosaic has been emulated using ACS $F814W$ images[1] acquired by the *Hubble* Space Telescope (HST, Koekemoer et al. 2007; Massey et al. 2010). It has been generated by re-binning and smoothing the HST images to the *Euclid* pixel scale and resolution ($0.''1\,\mathrm{px}^{-1}$), and adding random Gaussian noise to match the planned *Euclid* VIS depth. The zero-point determination is described in Bohlin (2016).

Both the scientific image and rms map have been created using dedicated simulation software (see Appendix A for more information). Although the VIS-like image has the required depth and resolution, the $F814W$ filter ($0.7-0.95\,\mu$m) is narrower than the VIS one ($0.54-0.92\,\mu$m).

2.1.2. EXT- like images

The EXT-like ground-based data set in the g, r, i, and z bands is composed of coadded images from publicly available data in the COSMOS field, obtained by a Dark Energy Camera (DECam) community program[2]. The mosaics were created using the Cosmology Data Management system (`CosmoDM`, Mohr et al. 2012; Desai et al. 2012, 2015; Hennig et al. 2017).

The data processing and calibration follow the standard procedure, as outlined in Hennig et al. (2017), where the single-epoch images are astrometrically calibrated to 2MASS (Skrutskie et al. 2006) and internally photometrically calibrated using stellar sources that are common between pairs of overlapping images. Masking of transient artifacts is applied using the method described in Desai et al. (2016). The images are coadded and resampled onto the *Euclid* pixel grid ($0.''1\,\mathrm{px}^{-1}$). Table 1 lists the properties of the DECam coadded images prepared for this work.

2.1.3. NIR-like images

The NIR images (Y, J, and H) were produced by Terapix as part of the UltraVISTA release 1[3] (McCracken et al. 2012), and were resampled onto the *Euclid* pixel grid. These images have similar depths to those quoted in the *Euclid* Red Book (Laureijs et al. 2011), so it was decided not to add any additional noise. It must be noted, however, that the Y, J, and H filters differ significantly from the equivalent *Euclid* filters since the *Euclid* ones are designed to leave no gap between the filters and the *Euclid* H band extends up to $2\,\mu$m. Table 1 shows the properties of the NIR-like coadded images. More details on the UltraVISTA images can be found in McCracken et al. (2012).

2.2. Photometry

Source detection was performed on the VIS-like image. The PSF of all images were homogenized to the g-band one, which has the poorest resolution among the eight images. The fluxes were extracted from the images using `SExtractor` 2.19.5 (Bertin & Arnouts 1996) in dual-image mode. Total flux measurements were performed on the VIS-like image from `SExtractor` `FLUX_AUTO` counts. Fluxes in the other bands were measured in apertures on PSF-matched images. The conversion between counts in band X, C_X, as measured by `SExtractor`, and fluxes, F_X, in μJy was performed using

$$F_X = C_X\, 10^{0.4(23.9-\mathrm{ZP}_X)}. \quad (1)$$

The zero-points (ZP_X) of band X can be found in Table 1.

Aperture fluxes on the PSF-matched images were computed in circular apertures of n times the flux profile full width at half maximum (FWHM) of the PSF in the g-band image. The flux in each band was scaled to total flux according to the following equation:

$$F_{X,\mathrm{tot}} = \left(\frac{F_{X,\mathrm{aper}}}{F_{\mathrm{VIS,aper}}}\right) F_{\mathrm{VIS,tot}}, \quad (2)$$

where $F_{X,\mathrm{aper}}$ is the measured aperture flux in band X for which the PSF has been matched to the one of the g-band, $F_{\mathrm{VIS,aper}}$ is the aperture flux in the VIS-like-band with PSF matched to the g-band one, and $F_{\mathrm{VIS,tot}}$ is the total flux extracted in the VIS-like-band with its original PSF. Fluxes were obtained for three different apertures sizes, with $n = 1, 2$, and 3 times the FWHM of the g-band PSF. Flux uncertainties were also scaled on the basis of the ratio between the total VIS-like flux and the VIS-like PSF-matched one measured in an aperture:

$$F_{\mathrm{err},X,\mathrm{tot}} = (F_{\mathrm{err},X,\mathrm{aper}}) \frac{F_{\mathrm{VIS,tot}}}{(F_{\mathrm{VIS,aper}})}, \quad (3)$$

where $F_{\mathrm{err},X,\mathrm{aper}}$ is the aperture flux error in band X.

To take into account pixel correlations coming from the resampling, which underestimate the measured flux errors, the errors were corrected according to the difference measured in 2″ diameter apertures between the sky background noise and the mean variance computed from the weight maps. The corrections (multiplicative factors on flux errors) are given in Table 1.

[1] http://irsa.ipac.caltech.edu/data/COSMOS/images/acs_mosaic_2.0/

[2] Data available on http://archive1.dm.noao.edu/; program number: 2013A-0351.

[3] http://www.eso.org/sci/observing/phase3/data_releases/ultravista_dr1.html

Table 1. Properties of the Ext-like, NIR-like, and VIS-like images used to generate the "*Euclid*" mosaics.

	PSF-*FWHM* [arcsec]	Depth [AB mag, 10σ]	Native pixel scale [arcsec pixel^{-1}]	Zero-point [AB mag]	Error correction factor
g	1.250	24.20	0.27	31.90	1.247
r	1.151	23.85	0.27	32.32	1.259
i	1.005	22.96	0.27	30.19	1.380
z	0.807	22.45	0.27	31.26	1.191
Y	0.855	23.81	0.15	30.00	2.884
J	0.831	23.59	0.15	30.00	2.582
H	0.800	23.13	0.15	30.00	2.377
VIS-like	0.200	24.50	0.03	25.49	1.038

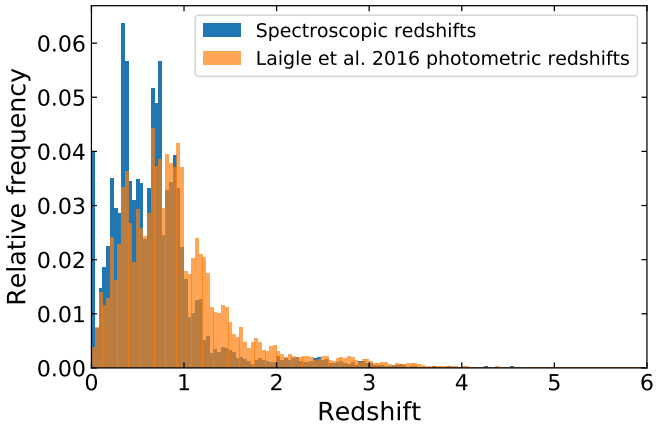

Fig. 2. Distribution of the full sample of reliable redshifts. Spectroscopic-redshifts are in blue (27 872 sources), while Laigle et al. (2016) 30-band photo-zs are in orange (107 267 sources).

2.3. Catalog

All objects detected by `SExtractor` are included in the catalog. Objects with problematic photometry (mostly located at the borders of masked regions), bad `SExtractor` flags, or with zero weight in at least in one of the weight maps (mostly objects outside the near-IR footprint) have been flagged. Galactic extinction correction factors for the fluxes in all bands, derived from Schlegel et al. (1998), are also provided for all sources but were not applied directly to the extracted photometry.

The catalog is divided into two regions based on right ascension α, defining two sub-catalogs: the calibration catalog, with $\alpha > 150.^{\circ}125$; and the validation catalog, with $\alpha \leq 150.^{\circ}125$. The first catalog is used for the calibration of the different methods to be tested, and the second one is used to assess the performance of all the codes. The number of sources is 198 435 in the calibration sample and 192 864 in the validation sample.

Both photometric catalogs have been matched to the master spectroscopic catalog maintained by M. Salvato, which is available within the COSMOS collaboration and contains approximately 50 000 objects (including around 30 000 with high-confidence flags), which serves as our primary reference to measure photo-z performance. Only the spec-zs for the calibration sample are provided as part of the challenge.

In addition to spectroscopic-redshifts, we matched the photometric catalog with the highly reliable photo-zs from Laigle et al. (2016, hereafter L15) that have been obtained using deep, 30-band photometry (scatter $\sigma = 0.01$ and outlier fraction $\eta = 1.7\%$ for $22 < i_{\rm AB} < 23$ sources). Figure 2 compares the distribution in redshift of the spec-zs and of the L15 photo-zs. The 30-band photo-zs are also included in the calibration sample and can be used to calibrate the different methods.

Stars and active galactic nuclei (AGN) in the field are separated from the galaxies by matching our catalog to the point source catalog from Leauthaud et al. (2007), the L15 catalog, and the catalog of X-ray detected AGN from Marchesi et al. (2016). Objects are classified as stellar if they are flagged as such in Leauthaud et al. (2007), or when the spec-zs or the L15 photo-zs are consistent with 0 and the `SExtractor FLUX_RADIUS_DETECT` measurements are smaller than 1.5 pixel. Sources with X-ray detections are flagged as AGNs.

2.4. Euclid shear sample

Unless specified, we focus on the sources in the *Euclid* shear sample. This sample is defined by the set of galaxies with a detection in the VIS-like band with signal-to-noise ratio $S/N > 10$, with $m_{\rm VIS} < 24.5$, that are not flagged as having poor photometry, and that are not flagged as AGNs. In addition, galaxies in the *Euclid* shear sample must have a photo-z in the range $0.2 < z_{\rm phot} \leq 2.6$, meaning that the detailed composition of this sample is method dependent.

3. Methods

Thirteen different methods have been tested on the data set (see Table 2 for a summary). In this section we provide brief descriptions of the algorithms and of the configurations that were used. A requirement of the challenge is that all methods should provide photo-z point estimates (a single value representative of the PDZ, e.g., the mean, the median, the mode, etc.), a probability distribution of the redshift (PDZ), as well as a usability flag (USE flag equals to 0 or 1) for all sources in the validation catalog. This usability flag, indicating whether the participant considers the photo-z estimate reliable or not, is defined freely by the participants. The results for rejected objects with USE = 0 are not accounted in the computation of the metrics. In the spirit of the challenge, the choices for the configuration of the different methods using the calibration catalog data were made independently by the subgroups of authors that ran the codes.

3.1. `LePhare`

`LePhare` (Arnouts et al. 2002; Ilbert et al. 2006) is a template-fitting method. The photo-zs for this work were derived following the recipes outlined in Ilbert et al. (2009). Thirty-three spectral energy distribution (SED) templates were used: the 31

Table 2. Summary of the different methods compared in this work.

	Type	Rejection
Le Phare	Template-fitting	Weak
CPz	Random forest classification + template-fitting	Weak
Phosphoros	Template-fitting	No
EAzY	Template-fitting	Strong
METAPHOR	Machine-learning: neural network	Strong
ANNz	Machine-learning: neural network	No
GPz	Machine-learning: Gaussian processes	Weak
GBRT	Machine-learning: boosted decision trees	Weak
RF	Machine-learning: random forest	No
Adaboost	Machine-learning: boosted decision trees	No
DNF	Machine-learning: nearest neighbor	Strong
frankenz	Machine-learning: nearest neighbor	Strong
NNPZ	Machine-learning: nearest neighbor	No

Notes. Columns are: the name of the code, the type of the approach (template-fitting or machine-learning) and whether a rejection of the results is applied or not. We qualify as strong a rejection of more than 15% of the full spectroscopic sample (more than 10 594 sources remaining; see Table B.1), otherwise it is considered to be a weak one.

COSMOS templates (Ilbert et al. 2009), that includes elliptical and spiral galaxies from the Polletta et al. (2007) library (some of them being linearly interpolated to refine the sampling in color-redshift space) and young and blue star-forming galaxies whose templates were generated with Bruzual & Charlot (2003) stellar population synthesis models; and two templates of elliptical galaxies generated with an exponentially decaying star-formation history (SFH; following Ilbert et al. 2013). Extinction was added as a free parameter ($E_{B-V} < 0.5$) on templates of type Sc and bluer. The Calzetti et al. (2000) attenuation curves were considered, adding a possible bump at 2175 Å, as well as the Prevot et al. (1984) attenuation curve. Emission lines were added to the templates using an empirical relation between UV light and emission line fluxes (Ilbert et al. 2009). Line fluxes were allowed to vary by a factor 2, but without changing the emission line ratios.

In the fit of the 2-FWHM photometry, a minimum error of 0.01 mag for all the visible bands was applied, and a minimum error of 0.03 mag for all the NIR *Euclid* bands was applied. A cut in absolute magnitude was applied, discarding all solutions with galaxies brighter than $M_g = -24$. An optimization of the zero-points was made using the method of Ilbert et al. (2006). Offsets as large as 0.07 mag were applied to two bands, namely the i and Y bands.

Redshift point estimates are given as the median of the marginalized PDZ. All sources with a 68% confidence intervals around the median larger than $0.3(1 + z)$ were flagged with USE = 0.

3.2. CPz

Classification-aided photometric-redshift estimation (CPz; Fotopoulou & Paltani 2018) is a hybrid approach to compute photometric redshifts. This methods uses random forest (Breiman 2001) to assign each object its optimal class, and then uses traditional SED fitting (Le Phare; Arnouts et al. 1999) for the photometric-redshift estimations. The goal of this method is to use a restricted library of templates optimized for each of the galaxy classes considered, aiming to reduce degeneracies between models. It can be considered as a generalization of the approach described in Salvato et al. (2011).

The data were split into three equal parts, used for training, validation, and testing, respectively. Three distinct random forest classifiers were trained to assign each object into: (i) star versus not star; (ii) one of the five galaxy classes (passive, starforming, starburst, AGN, or QSO); and (iii) photometric redshift outlier. All models were fit to the data and labels were assigned based on the SEDs that provide the best photometric redshifts. The galaxy models (passive, starforming, starburst) are the 31 COSMOS templates used in Ilbert et al. (2009), while the AGN and QSO templates are from Salvato et al. (2011). A detailed description of the model set up can be found in (Fotopoulou & Paltani 2018, Case III). Briefly, models were generated at $0 < z < 6$ with $\Delta z = 0.01$. Attenuation values $E_{B-V} = 0$, 0.05, 0.1, 0.15, 0.2, and 0.3 were used. Emission lines were added only for the normal galaxy templates, as the AGN and QSO templates are empirical and already contain emission lines.

The classification was performed in color space, by taking all color combinations of the input photometry without any input redshift information. Once the three classifiers were trained and applied to the entire sample, the redshift solution was assigned using the model library identified by the classifier as optimal. Additionally, sources classified as stars ($P_{\rm star} \geq 0.5$) or outliers ($P_{\rm outlier} \geq 0.5$) were rejected (USE = 0). Since this application concerns the estimation of photometric redshifts for *Euclid*, sources that are classified as AGN or QSO are also rejected, since they typically have lower quality photo-z.

3.3. Phosphoros

Phosphoros (Paltani et al., in prep.) is a Bayesian template-fitting tool developed with the aim of being run in a computer-intensive processing environment while including most of the advanced features found in similar codes, such as the use of upper limits, zero-point corrections, consideration of emission lines, various intrinsic extinction curves, etc. Phosphoros will implement unique features, like complex user-defined priors (e.g., from luminosity functions), the choice between different intergalactic medium prescriptions, the sampling of the posterior, etc. Because it is still under active development, the only advanced and unique feature that we use here is the improved treatment of Galactic reddening (Galametz et al. 2017).

The 2-FWHM aperture photometry was selected for the EXT-like and NIR-like bands, and the total flux for the VIS-like band. The photometric data were fit with the 31 COSMOS galaxy (SED) templates (see Sect. 3.1) with a similar configuration as in Ilbert et al. (2013), from $z = 0.01$ to $z = 5.99$ with step size of $\Delta z = 0.02$. Intrinsic reddening was set as a free parameter, with $E_{B-V} \leq 0.5$ and several extinction laws (Prevot et al. 1984; Calzetti et al. 2000 and modified Calzetti laws including a bump at 2175 Å as in Ilbert et al. 2009). For templates representing galaxies with types earlier than Sc, no extinction was added. The Hα to Hδ, [O II] 3727 Å and [O III] 4959+5007 Å emission lines were added to all templates using an empirical relation between Hα and other emission line fluxes, which were recalibrated using line fluxes measured from the Sloan Digital Sky Survey (Thomas et al. 2013). The Milky Way reddening was treated as prescribed in Galametz et al. (2017) by applying a reddening correction to the templates and fitting uncorrected photometry. Zero-point corrections to the photometric calibration were computed in the same way as in Ilbert et al. (2006) using 2000 randomly selected galaxies with spec-zs from the calibration catalog. No luminosity prior has been used.

The PDZs are constructed by marginalizing the likelihood over the template and reddening dimensions. The point estimates used in the rest of the analysis were computed from the mode of the PDZ for each object. Finally, no rejection was made on the quality of the results (USE flag was set to 1 for all sources).

3.4. EAzY

The setup for the `EAzY` code (Brammer et al. 2008) was kept close to the default configuration. All template combinations of the seven base SED components with added emission lines were allowed, plus a young, heavily dust-reddened galaxy SED (which is not allowed to be combined with the other SEDs). The extended r-band magnitude based prior, $p(z|m_r)$, was applied and a flat ΛCDM cosmology with $H_0 = 70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and $\Omega_\mathrm{m} = 0.3$ for luminosity computation was used. The run was performed using the 2-FWHM aperture photometry without the VIS-like-band. A single best χ^2 value among the possible template combinations was returned at each redshift, which was then combined with the magnitude-based prior to produce the galaxy PDZs.

Similar to other template-fitting based methods, the `EAzY` code includes the flexibility to apply corrections to photometric zero-points and a systematic uncertainty in measured photometric fluxes. However, there is also a wavelength-dependent template uncertainty function which is controlled by a parameter that governs its amplitude. These nine parameters (namely seven zero-points, fractional systematic flux error and template error function amplitude) were optimized using the spectroscopic training sample and the Python function `minimize` from the `scipy.optimise` package. Values were initialized at zero for the zero-point adjustments, 3% for the systematic flux error, and 0.7 for the amplitude of the template error function. The loss function is a linear combination of the normalized median absolute deviation, mean point redshift bias, Kullback–Leibler divergence of the histogram of probability integral transform values (see Sect. 4.2 for more information), and outlier fraction. Each term in the loss function was scaled such that a value of unity represented good performance. The point redshift used for the first two terms, and for the tomographic bin assignment, is `z_peak`, the mean redshift of the most probable peak in the PDZ. Finally, objects were flagged as unreliable if their odds value was smaller than 0.91. The odds value quantifies the extent to which a PDZ is single-peaked (see Brammer et al. 2008), and this value was chosen as a compromise between sample completeness and performance on the same set of metrics that were used in the loss function.

3.5. METAPHOR

`METAPHOR` (Machine-learning Estimation Tool for Accurate PHOtometric Redshifts; Amaro et al. 2019; Cavuoti et al. 2017) has a modular workflow, designed to produce the redshift point estimations and the PDZs. Its internal photo-z estimation engine is based on the MLPQNA machine-learning model (Multi Layer Perceptron with Quasi Newton Algorithm; Brescia et al. 2013; Cavuoti et al. 2015).

The key concept of `METAPHOR` is to perform a series of independent photometry perturbations to take into account the contribution of the uncertainty induced by the photometric errors within the PDZs. In other words, the idea is to obtain an estimation of the photo-z PDZs based on the predictive performance evaluation of the trained MLPQNA model by varying the magnitudes within the photometric errors and considering the distribution of the multiple output as the PDZ. The perturbation method is based on the addition of a variable random Gaussian noise to the photometry and a polynomial fitting of the photometric trend to reproduce the inner distribution of the error.

In practice, each PDZ was based on the following steps: (i) training of MLPQNA with unperturbed SEDs (the training set); (ii) producing N different instances of any source SED (the blind test set) contaminated by photometric noise; (iii) deriving $N + 1$ photo-z estimates for the sources with the trained model (i.e., N perturbed + the original one); and (iv) binning in photo-z of the $N + 1$ values, thus calculating for each one the probability that a given photo-z estimation belongs to each bin (i.e., obtaining the PDZ). In the particular case of this *Euclid* challenge, $N = 999$ was used. The point estimate is the value among the $N + 1$ values that is the closest to the non-perturbed value.

The training was done with the 2-FWHM photometry in all bands, considering all galaxies (including AGNs) with spec-zs and flagged as having proper photometry. In order to introduce a quality flag for the estimates, a two-step analysis was performed. First, a selection on the photometry in which all objects with $S/N \leq 3$ in any of the $griz$ bands, $S/N \leq 5$ in any of the YJH and VIS-like bands, or a `SExtractor` detection flag $\geq$4, were marked with the flag USE = 0. Second, a further refinement of the flag assignment was performed through a selection on the PDZ to avoid overly wide PDZs. The criteria were: the maximum value of a PDZ must be $\geq$0.09; the width of its primary peak $\leq$0.44 in redshift; and the overall distribution must be $\leq$2.

3.6. ANNz

`ANNz` (Collister & Lahav 2004) is a neural-network-based photometric redshift code that uses a training set with both photometric and spectroscopic information to learn the mapping between the color-magnitude space of galaxies to their redshifts. The learning algorithm minimizes the mean-squared error between predicted and (assumed to be) true, spectroscopic-redshifts. The learned function is then an estimate of the mean of the conditional distribution $p(z|\boldsymbol{f})$, where z denotes the redshift and $\boldsymbol{f}$ the

vector of galaxy colors and their magnitudes. The learned model is then applied to the full data sample to obtain photometric redshift estimates.

To obtain error bars on these point estimates, one can subsequently train an additional neural network to predict the mean-squared error between true (i.e., spectroscopic) redshift and the predicted redshift from the previously trained model. This is done using the same basic setup, meaning, again by minimizing the mean-squared error. The resulting predictions from this second run then provide an estimate for the variance of the conditional distribution $p(z|\boldsymbol{f})$. These error bars can only be expected to be well calibrated if enough training data are available, the training data are representative, and the conditional distributions $p(z|\boldsymbol{f})$ are close to Gaussian. The interested reader is referred to Rau et al. (2015) for a discussion of the impact of these distributional assumptions on photo-z results.

The calibration sample was split into two representative subsamples. The first one was used for training, and the second one was used for testing the models.

3.7. GPz

GPz is a machine-learning tool that models the relation between input data (e.g., observed magnitudes, which we call "color" for simplicity) and an output value (the redshift). The model used by GPz is a linear combination of multivariate Gaussian functions (called "basis functions"; here 100 are used). In addition to learning the mean relation between colors and redshifts, GPz also learns the scatter of the redshift at a given position in color space, as well as the density of the training data. It uses this information, together with knowledge of the uncertainties on the observed colors, to make a prediction of the PDZ. At a given position in color space, the predicted PDZ will be broader if: (i) the colors are uncertain; (ii) the range of matching spec-z is large; or (iii) there is a lack of training data. A limitation of this model is that the distribution is forced to be Gaussian (Almosallam et al. 2016a,b). All the predictions here are produced with the C++ version of GPz, available in the *Euclid* GitLab as "PHZ_GPz". Here, the model was trained on the shear sample, since this is the sample for which the metrics need to be optimized in *Euclid*. The 2-FWHM fluxes were used, and the fluxes were converted to "luptitudes" (Lupton et al. 1999) before prediction and training.

First, GPz models the distribution in color space of the validation data set (the one for which we want to make predictions) using Gaussian mixture models. It then applies this Gaussian mixture to the training set to: (i) weight the training data so its color distributions match the validation data; and (ii) split the color space in several (typically 5) distinct regions in which separate GPz models will be trained. The first point deals with any potential bias in the color distribution of the training set, while the second point effectively increases the number of basis functions used to model a given region of color space without paying for the full computational costs.

3.8. Gradient boosted regression trees

Gradient boosted regression trees (GBRT) is a machine-learning method based on the sci-kit learn gradient boosted decision tree algorithm (Friedman 2001; Pedregosa et al. 2011). For its training, galaxies and AGNs with good quality spectroscopic or 30-band photometric redshifts from L15 and detected in at least four bands were selected from the calibration catalog, leading to a training sample of around 1.4×10^5 sources. This sample size is increased by an order of magnitude by synthesizing 10 brighter and fainter versions of each source. The 2-FWHM aperture photometry in the g, r, i, and z bands, the 1-FWHM aperture photometry in the Y, J, and H bands, and the VIS total photometry were used to train the algorithm.

A point estimate was determined for each source, and a PDZ was constructed through processing of 1000 realizations perturbed by a Gaussian error for each source. GBRT provides an indication of the most useful bands for the photo-z determination, those being the g, Y, J, H, and VIS-like bands. Sources located in regions of this color space that were not covered by sources from the training sample were rejected (USE = 0).

3.9. Primal Random Forest

The Primal Random Forest (RF) is based on the sci-kit learn random forest regressor (Breiman 2001; Pedregosa et al. 2011) wrapped in the Primal framework[4]. The training was done by selecting all sources with reliable spectroscopic-redshifts in the calibration catalog. The features used were the 2-FWHM aperture fluxes in all standard bands and the total fluxes in the VIS-like band, along with the flux ratios and flux errors. The calibration sample was split into training (20%) and testing (80%) sets using a reshuffling procedure with stratified sampling to insure that both sets were representative of the full sample. RF is optimized by performing a recursive feature elimination, selecting the most important features that provide the minimum outlier fraction.

The validation set was processed 5000 times with perturbed fluxes according to their errors. The PDZs were constructed by binning the 5000 results for each source. The point estimates are the modes of the constructed PDZs. No rejection was made on the quality of the results, so that the USE flag is set to 1 for all objects with good photometric flags.

3.10. Primal Adaboost

The Primal Adaboost method is the sci-kit learn Adaboost regressor algorithm (Freund & Schapire 1997) wrapped in the Primal framework. We used boosted decision tree regressors. The training and the processing were done in the exact same way as for the RF, which is described in Sect. 3.9.

3.11. DNF

DNF (Directional Neighborhood Fitting; De Vicente et al. 2016) computes the photo-z of a galaxy by a linear combination of multi-band fluxes. The coefficients of the prediction hyperplane are determined by fitting the equation with a subsample of neighbors within a reference sample whose spectroscopic-redshifts are known. A novel metric ("directional neighborhood") is defined to account simultaneously for the magnitudes and colors of the galaxies. The PDZs are computed from the residuals of the fit and reflect the uncertainties and degeneracies associated with individual photo-z predictions (see details in De Vicente et al. 2016). DNF also produces a second photo-z ($z_{\rm phot,2}$ estimate) as the redshift of the nearest directional neighbor in the reference sample. The stacking of $z_{\rm phot,2}$ values for the whole sample provides a reference redshift distribution estimation, if the target galaxies are well represented within the training sample.

[4] https://github.com/andreatramacere/primal

DNF was run on galaxies only, using the 3-FWHM photometry. DNF provides an error estimation of individual photo-zs that accounts for flux uncertainty, also tagging the lack of neighbor reference samples. This parameter allows one to cut the samples according to different precision, bias, or completeness requirements. In this test, precision was prioritized over bias and completeness, producing an aggressive cut of 50% of the sample. Other configurations are possible such as those focusing only on removing the most unreliable photo-zs.

3.12. frankenz

frankenz (Tanaka et al. 2018; Speagle et al., in prep.) adopts a Bayesian-oriented nearest-neighbors-based approach that attempts to properly account for measurement errors within both training and testing sets when making photo-z predictions. Neighbors were selected using a Monte Carlo approach over repeated realizations of the photometric errors, after which priors over the training set (here assumed to be uniform) and the likelihoods between each unique training-testing object pair were computed explicitly in flux space. PDZs were then constructed using a posterior-weighted average of each object's redshift kernel. Objects with large best-fit reduced χ^2 values among the set of nearest neighbors were flagged not to be used (USE = 0).

3.13. NNPZ

NNPZ (Nearest-Neighbor Photometric Redshift) is a machine-learning algorithm that consists in a k-nearest neighbor method in flux space, developed by J. Coupon, that is designed to produce PDZs and was applied to the HSC-SSP survey (Tanaka et al. 2018).

An improved version of the algorithm was used here, which takes into account errors when searching for the neighbors and weights them according to some distance definition. For efficiency, in this implementation of NNPZ the process is split into three stages. First, NNPZ reduces the search space by selecting a candidate set of neighbors using a *k-dimensional tree* and Euclidean distances, which allows for look-ups in $O(\log n)$ steps. Over this initial candidate set, the final neighbors are searched using a χ^2 distance, which takes into account both the errors of the reference and the target object. Finally, the weights are computed using the likelihood of the χ^2.

The training was done using the Galactic-reddening corrected 2-FWHM aperture photometry of the sources that were not flagged as stars or AGN. The labels were the reliable L15 photo-zs, restricted to $0 < z \leq 6$. For the first stage, NNPZ selected 2000 neighbors using the Euclidean distance, then later reduced their number to 30 using the χ^2 distance. The PDZs were constructed by combining the L15 PDZs of the weighted neighbors.

The point estimate is the mode of the PDZs for each source. No rejection was made on the quality of the results, so that the USE flag was set to 1 for all objects with good photometry flags.

4. Results

In the following, we consider the *Euclid* shear sample (see Sect. 2.4). In the *Euclid* context we focus on the performance of the different methods in the $0.2 < z < 2.6$ photo-z range and for source with USE = 1. In the rest of the analysis, we refer to this selection as the *Euclid* selection. We point out that the *Euclid* selection is different for each method, since each method assigns different photo-zs and has different flagging schemes.

4.1. Point estimates

First, we look at the point estimates and assess the quality of the results through the following commonly used metrics: the normalized median absolute deviation of the residuals

$$\sigma = 1.4826 \times \text{median}\left(|\,\Delta z - \text{median}\,(\Delta z)\,|\right),$$

where $\Delta z = (z_{\rm spec} - z_{\rm phot})/(1+z_{\rm spec})$ is the scaled residual between the photo-z and the reference redshift; and the fraction η of outlier sources for which $|\Delta z| > 0.15$.

Figure 3 shows the density map of the photo-z point estimates versus the spec-zs for all thirteen methods. The same plots using the 30-bands photo-zs as reference redshift can be found Fig. C.1. All sources without photo-zs are set to $z_{\rm phot} = 0$, which explains the horizontal lines in some of the plots, and are treated as outliers in the computation of the metrics. METAPHOR shows a systematic photo-z at $z = 4.12$, which corresponds to the highest redshift in the training sample they considered. The statistics associated with the plots are presented in a graphical form in Fig. 4, and all the values are provided in tables in Appendix B. We note some results that appear similar in Fig. 3, like those of Phosphoros and CPz; this is due to the similarity of the approaches (template-fitting) and configuration (31 COSMOS templates from Ilbert et al. 2013), even if the codes are different. On the other hand, the difference in the results between LePhare and CPz can be explained by the differences in the definition of the point estimate, being the median of the PDZ for LePhare and the mode for CPz, even if CPz uses LePhare for the fitting of the templates. Further tests have shown that when run in identical configuration, template-fitting methods provide identical results. This means that the differences observed in the results are not due to differences in performance of the template-fitting methods, but rather to variations in their configurations.

Figure 4 shows the metrics associated with different reference redshifts and selections applied to the sources. In the top left panel, $\sigma_{\rm all}$ and $\eta_{\rm all}$ are plotted against each other for the total spectroscopic sample (12 463 sources with highly reliable spec-z measurements). With its large outlier fraction, frankenz differs greatly from the rest of the methods in the plot. This is due to the sources for which no photo-z are provided, visible in Fig. 3 with $z_{\rm phot} = 0$. Machine-learning methods seem generally to perform better than the template-fitting ones, especially Adaboost or ANNz. The top right panel of Fig. 4 presents metrics for the spectroscopic sample, but only considering sources with USE flag equal to 1. In this case, we see some improvement in the results of the methods that apply rejection of the sources for which the predictions are considered less reliable. This phenomenon is particularly obvious for METAPHOR, which shows the best results after this rejection. This demonstrates that the USE flags are able to correctly identify a good fraction of the incorrect predictions, and that they enhance the precision of the results, at the expense of completeness.

For the *Euclid* selection, $\sigma_{\rm Euclid}$ and $\eta_{\rm Euclid}$ are presented in Fig. 4 (bottom left panel). In this range of redshifts, the results are better for all the methods. Phosphoros and CPz show great improvements, since the selection removes low photo-z sources that are poorly constrained due to the absence of u-band fluxes in the data. Here again, METAPHOR presents the best values for these metrics.

In order to take into account the fact that the spec-z sample is not representative of the color space of all galaxies, we follow the

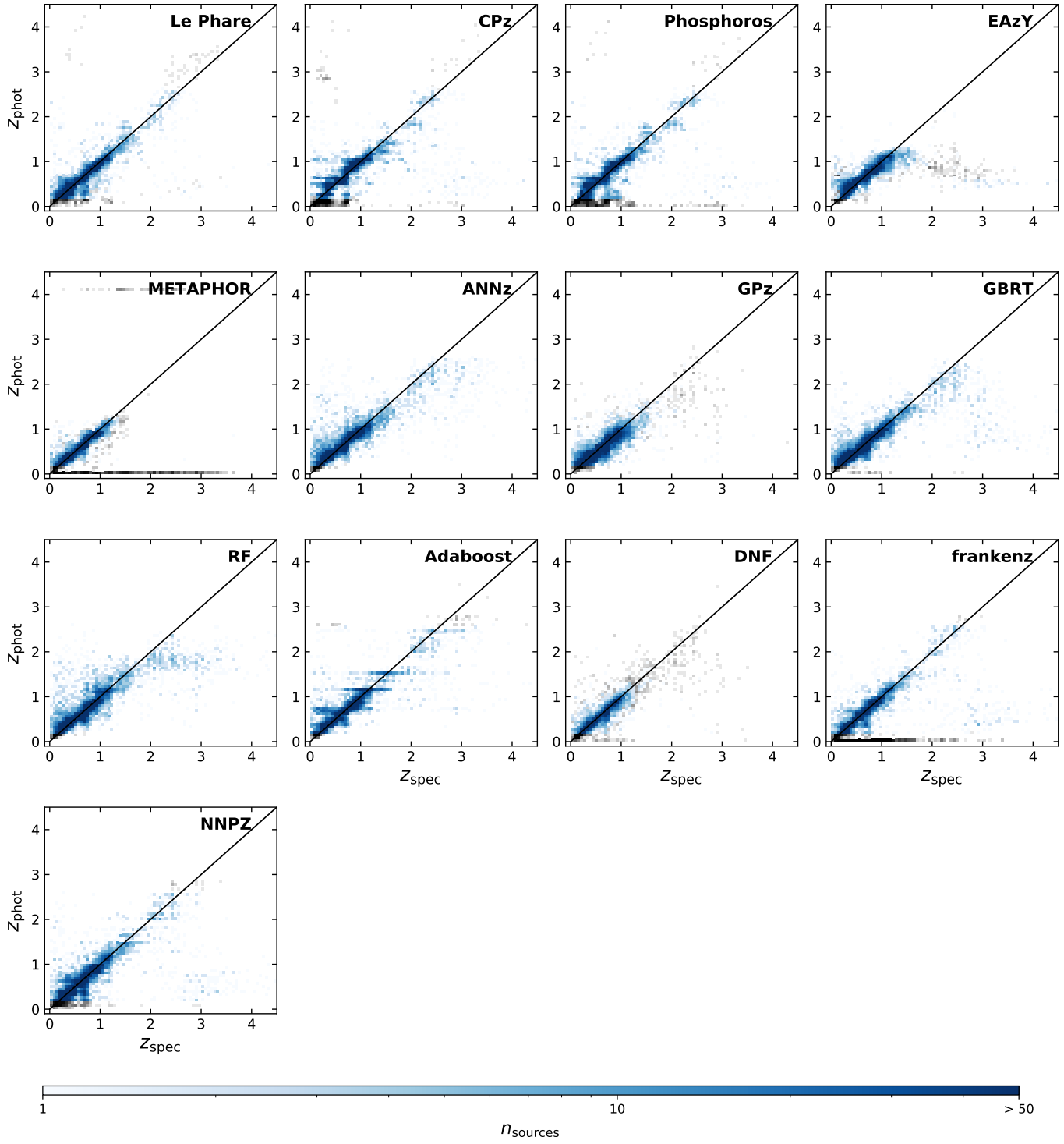

Fig. 3. Density maps of photo-z versus spec-z for all the tested methods: blue are sources within the *Euclid* sample; gray are sources outside of the *Euclid* sample. The statistics on the photo-zs are presented in Fig. 4 and listed in Tables B.1 and B.2. Undefined or negative photo-zs have been set to 0, explaining the presence of horizontal lines in some panels (e.g., `METAPHOR` and `frankenz`).

approach of Lima et al. (2008). We assign weights to the spectroscopic sample depending on the distances of the 100 nearest neighbors each object has in the color-mag$_{VIS}$ space of the full shear sample using a nearest-neighbor method. We compute the indicators $\sigma_{color-space}$ and $\eta_{color-space}$ with these weights, presented in the bottom right panel of Fig. 4. Both scatter and outlier fractions become poorer for most of the methods as one would expect, with the exception of `METAPHOR` that shows only a small reduction in performance. Summing the weights of the sources in the selection of each method ($N_{color-space}$ in Table B.3) and comparing this sum to the sum of the weights for all the sources that should be in the *Euclid* sample (9384.4) gives an estimate of the fraction of sources kept by the methods for the photometric sample. For `METAPHOR` the ratio between the two values is 1/3, and the ratio is 1/4 for `DNF` (the median value for all the methods is ∼0.86), meaning that the majority of the sources is rejected

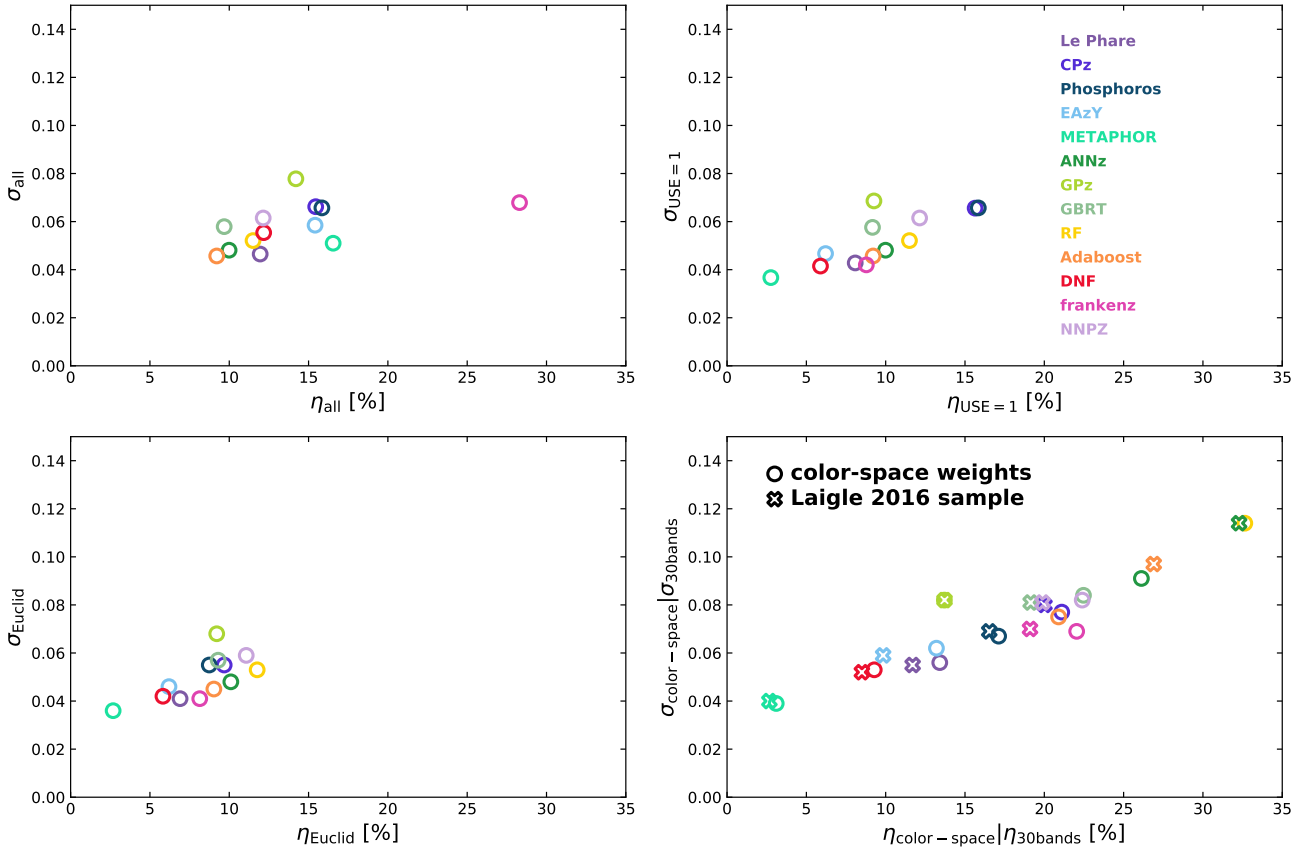

Fig. 4. Point estimates metrics results comparison for all the methods. Circles represent the spectroscopic sample and crosses are the L15 one. *Top left*: scatter (σ) versus outlier fraction (η) of all methods for the whole spectroscopic sample (12 463 sources). *Top right*: $\sigma_{\mathrm{USE=1}}$ versus $\eta_{\mathrm{USE=1}}$ for USE = 1 selected sources for each method (see Table B.1 for all the values). *Bottom left*: σ_{Euclid} versus η_{Euclid} for the *Euclid* selection (see Table B.2). *Bottom right*: $\sigma_{\mathrm{30\text{-}bands}}$ versus $\eta_{\mathrm{30\text{-}bands}}$ using the L15 photo-z as reference redshift plotted as crosses (see Table B.4 for all the values) and *Euclid* sample results weighted with the color-space weights to match the spectroscopic sample to the photometric one plotted as circles (see Table B.3). RF values are outside the limits of the plot for the L15 sample due to a large outlier fraction.

in the photometric sample in order to keep the precision of the photo-zs at the level of the spectroscopic sample.

Another estimate of the quality of the photo-zs over the full color space can be obtained by comparing our photo-zs with the 30-band ones of Laigle et al. (2016). The underlying assumption is that the latter photo-zs are much more precise than those computed here, thanks to the much deeper and better sampled photometric data. The bottom right panel of Fig. 4 shows the scatter ($\sigma_{\mathrm{30\text{-}bands}}$) and outlier fraction ($\eta_{\mathrm{30\text{-}bands}}$) for the *Euclid* selection, computed with L15 photo-zs as reference redshifts (see Table B.4 for all the values). We note that the results with the L15 photo-zs are comparable to the color-space-weighted ones. The good match between color-space-weighted results and L15 allows us to consider either of these methods to be good approximations of the scatter and outlier fraction of the photo-z methods over the full photometric sample. In the following, we use both the weighted spectroscopic sample and the L15 sample, since we want to assess the quality of the results over the whole color space. Using both samples allows us to consider different systematics in the comparison: the weighted spec-z sample has more reliable reference redshifts, but might not represent the full photometric sample, since some part of the color space might not be covered at all; and the L15 sample, while complete in color space, contains less precise redshifts, as well as some catastrophic failures, because it is based on 30-band photo-zs. Methods trained on the spectroscopic sample can be expected to perform better on the weighted spec-z sample, while methods training on L15 data (NNPZ and GBRT), as well as the template-fitting methods, especially if they use the same templates as in L15, might present better results on this sample.

4.2. PDZs

Each method provides PDZs for every source. Compared to the point estimates, PDZs include all the information about errors and possible degeneracies of the measurements. We assess here the quality of the PDZs provided by all the methods. We consider only the *Euclid* sample selection (see Sect. 4.1).

The metric we choose to assess the quality of the results is the one chosen to express the photo-z requirements of *Euclid*. The sources are first distributed in photo-z bins depending on their point estimates. In each bin, the source PDZs, $P(z)$, are shifted by the values of the source spec-zs, $P(z - z_{\mathrm{spec}})$, in order to have the probability of the spec-z at the origin. Then, all the shifted PDZs of each bin are stacked, using color-space weights for the spectroscopic sample (see Sect. 4.1) and without weight for the L15 sample. For a bin centered on a redshift z, we compute the fractions of the stacked PDZ enclosed in $\pm 0.05(1 + z)$ around its mode (F_{005}) and the one enclosed in $\pm 0.15(1 + z)$ around the mode (F_{015}). We note that integrating the stacked PDZs around the mode implies that it exists a method to correct their biases; the current baseline is to apply a calibration in color space using self-organizing maps (Masters et al. 2015).

The quantities F_{005} and F_{015} measure the compactness of the distribution and can be compared to the scatter and outlier fraction of the point estimates. An F_{005} larger than 68% is the equivalent of the scatter being smaller than $0.05(1 + z)$ when considering PDFs. Likewise an F_{015} larger than 90% corresponds to the fraction of outliers, which are objects with $|z_{\rm phot}-z_{\rm spec}| > 0.15(1+z)$, being smaller than 10%. Therefore F_{005} and F_{015} are the equivalent of the scatter and outlier fraction when dealing with PDZs, and the $F_{005} > 68\%$ and $F_{015} > 90\%$ values are the equivalent to the requirements presented in the *Euclid* Red Book (Laureijs et al. 2011) when dealing with PDZs instead of point estimates.

Figure 5 shows the weighted spectroscopic sample F_{005} and F_{015} fractions for all the tested methods in 12 photo-z bins of width 0.2, from $z = 0.2$ to $z = 2.6$, as well as the number of sources per bin. In the distribution of sources per bin we see that the methods using strong rejection scheme, like `METAPHOR` or `EAzY`, provide very few, if any, predictions above $z = 1$. The F_{005} plot shows that template-fitting results and machine-learning results have distinct behaviors. `Phosphoros`, `Le Phare`, and `CPz` present a global level around $F_{005} \simeq 0.4$ with a strong peak in F_{005} at $0.6 \lesssim z \lesssim 1.2$, and a small drop around $z = 1.7$. The F_{015} values of the template-fitting methods have roughly the same shape as the F_{005} ones, with a base level of around 0.7 and less pronounced peaks. Some machine-learning methods (e.g., `GBRT`, `GPz`, and `DNF`) show $F_{005} < 0.4$ everywhere, highlighting the difficulties of machine-learning algorithms in general in producing informative PDZs. However, other machine-learning methods (e.g., `ANNz`, `Adaboost`, `RF`, `METAPHOR`, `frankenz`, or `NNPZ`) produce good PDZs according to the F_{005} and F_{015} metrics, although they experience sharp drops of F_{005} and F_{015} above $z \simeq 1.3$. We note that the machine-learning methods that show good results perform generally better than the template-fitting ones in the first three redshift bins, with the exception of `METAPHOR`, which shows better results than all the other methods until the $z = 1.2{-}1.4$ bin, above which all sources are discarded. After that, the template-fitting methods show better results. We notice the same behavior in the results for the L15 sample in Fig. 6. In Fig. 6, we see that the drop in F_{005} around $z = 1.7$ for the template-fitting results disappeared, possibly because the L15 PDZs are computed with similar template-fitting algorithms. We also see that the distinction between the template-fitting and machine-learning results is larger, due to a general decrease in F_{005} for machine-learning approaches. The diminution of F_{005} and F_{015} for the L15 sample could be explained by the uncertainties of the L15 photo-zs, however template-fitting methods showing similar results in both weighted spectroscopic and L15 sample mitigates this possibility. The results on the L15 sample show the struggles of machine-learning methods to provide sensible results for a sample with a color space not matching the one they have been trained on.

The quality of the PDZs can be assessed using other metrics. We test the performance of the PDZ using probability integral transform plots (PIT plot, Dawid 1984; D'Isanto & Polsterer 2018). We compute the cumulative distribution function (CDF) at the true or L15 redshifts for all the sources i:

$$C_i \equiv \mathrm{CDF}_i(z_i) = \int_0^{z_i} \mathrm{PDZ}_i(z)\,\mathrm{d}z. \quad (4)$$

Figure 7 presents the histograms of C_i for each of the tested methods, for both the color-space weighted spectroscopic sample and the L15 sample. If the PDZs correctly represent the probability distribution of the sources, the histograms should be flat. Outlier sources that have their spec-zs in the outskirts of their PDZs have their CDFs close to 0 or to 1 and produce the peaks at the edge of most of the histograms in Fig. 7. U-shaped PIT plots, like those of `Phosphoros` or `CPz`, show that their PDZs are under-dispersed, meaning that they are in general too narrow. On the other hand, the PIT plots of `GPZ`, `ANNz`, or `GBRT` present a bump, indicating that the PDZs are over-dispersed, hence the PDZs are generally too broad. Biased PDZs produce PIT plots

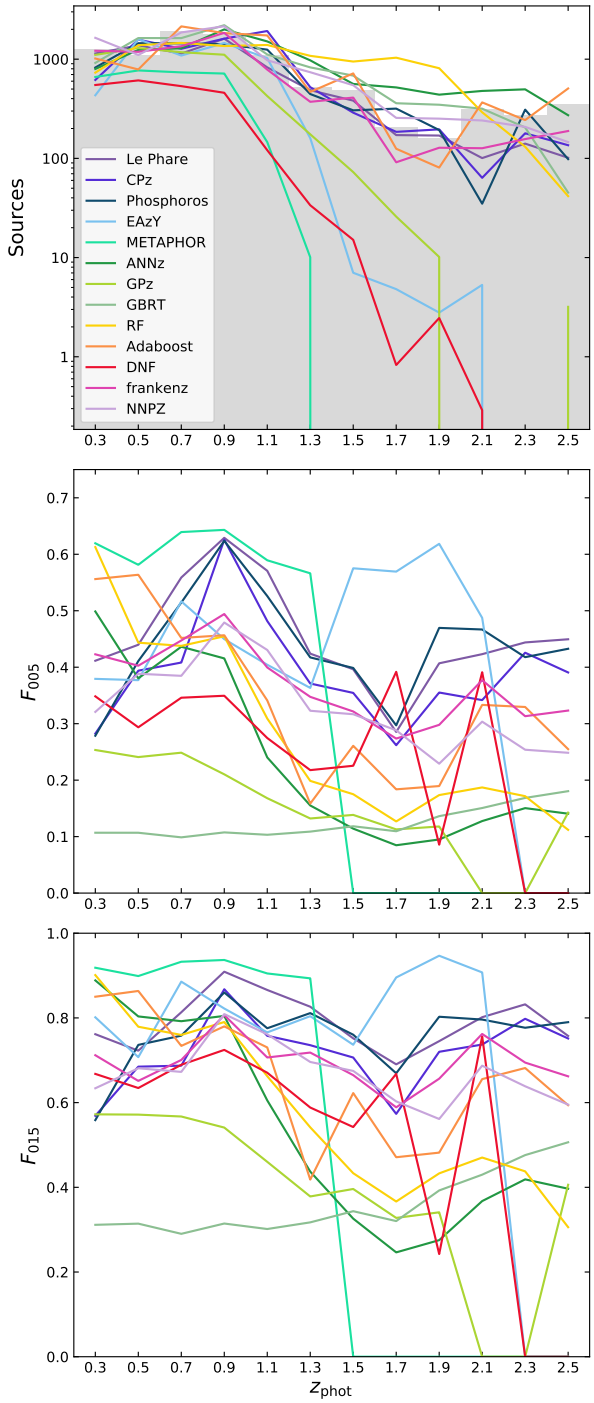

Fig. 5. PDZ metrics for the color-space weighted spectroscopic sample. *Top*: number of sources in the bin. The histogram of the distribution for the sources in the bins according to their spec-zs is shown in gray. *Middle*: fraction of the stacked-and-shifted PDZs in $0.05(1 + z)$ (F_{005}). *Bottom*: fraction of the stacked-and-shifted PDZs in $0.15(1 + z)$ (F_{015}) for all the tested methods. Fractions close to 1 in a bin indicate good results.

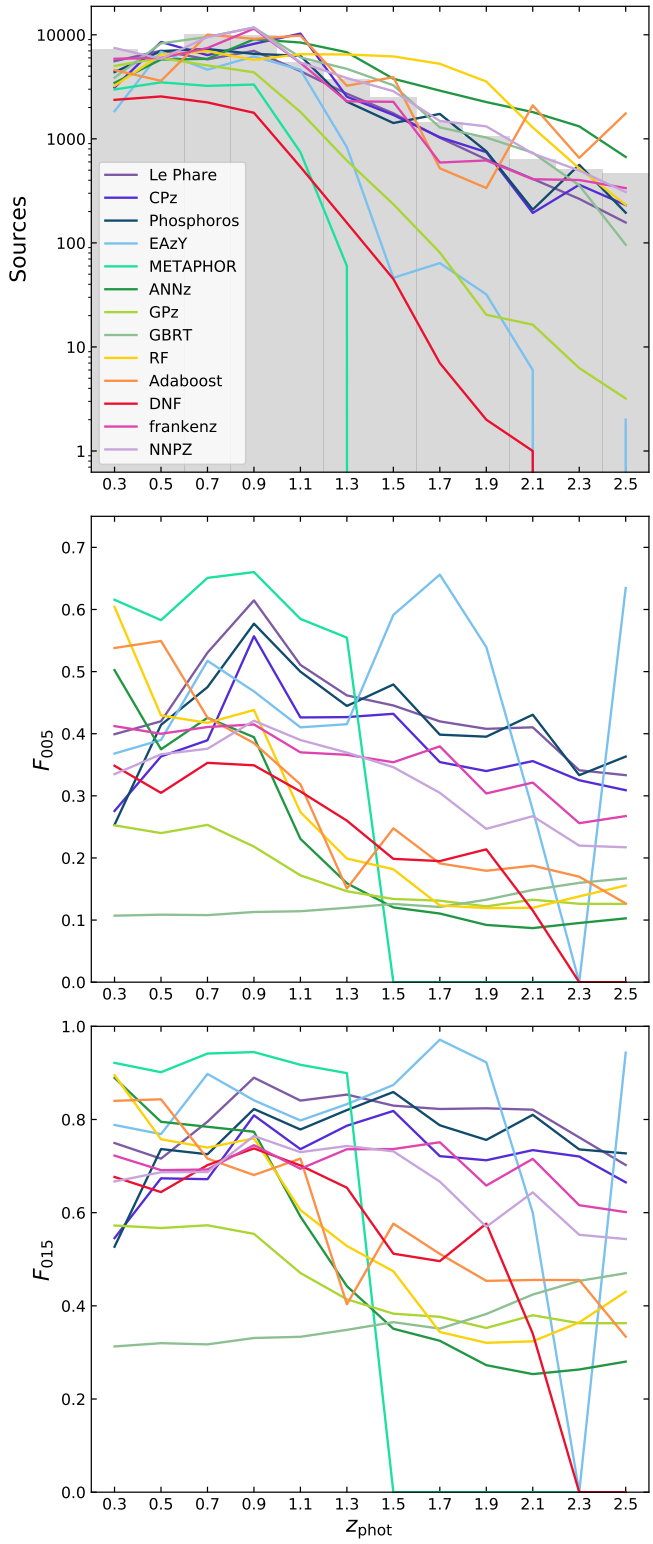

Fig. 6. Same as Fig. 5 for the L15 sample and the L15 photo-zs, instead of the spec-zs.

with a slope, like in the Le Phare, METAPHOR, DNF, or Frankenz histograms. We see that no method produces a perfectly flat PIT plot. We note also that there are no strong differences between the weighted spectroscopic sample and L15 sample PIT plots, with maybe the exceptions of frankenz and NNPZ that present flatter distributions for the L15 sample than on the weighted spectroscopic sample.

The other indicator often used to assess the quality of the PDZs is the continuous ranked probability score (CRPS, Hersbach 2000; D'Isanto & Polsterer 2018). It is defined as

$$\mathrm{CRPS}_i = \int_{-\infty}^{z_i} \mathrm{CDF}_i(z)^2 \, \mathrm{d}z + \int_{z_i}^{+\infty} [\mathrm{CDF}_i(z) - 1]^2 \, \mathrm{d}z, \quad (5)$$

and should be close to zero for a narrow PDZ at the true redshift. However, the CRPS would increase both in the cases of PDZ at the wrong redshift or a broad PDZ around the true redshift. The median CRPS (weighted in the case of the spectroscopic sample) provided in Table 3 give an indication of the overall quality of the PDZs for each method. We use the median instead of the mean due to the high CRPS values for some few outliers (more than 30 times the mean value in some cases, see Fig. D.1 for the complete distributions) increasing the mean value, which is thus less representative of the CRPS than the median for the majority of sources. Table 3 also reports the CRPSs obtained with precise (i.e., Dirac function) but biased PDZs or unbiased but dispersed PDZs (Gaussian with the true redshift as mean and a non-zero scatter), both tuned to provide values similar to those measured for the different methods. The CRPS is sensitive to both bias and scatter, with no possibility to distinguish between the two effects. CRPS values in Table 3 show that for the spectroscopic sample the majority of methods present a median CRPS of around 0.08. Some methods provide better results like METAPHOR, EAzY, or Le Phare with median CRPS values of around 0.04 to 0.06, and some methods have larger mean CRPS like ANNz, GPz, GBRT or RF, with median CRPS above 0.1, meaning they provide less sensible PDZs which can be also deduced from F_{005} and F_{015} indicators for GPz and GBRT. For the L15 sample, there are no strong changes from the results on the weighted spectroscopic sample.

5. Discussion

We have performed extensive tests of the performance of 13 photo-z algorithms using several metrics. One must keep in mind that all the analysis was done on the *Euclid* shear sample, which only contains galaxies in a restricted photo-z range of 0.2−2.6 (see Sect. 2.4). This means that our results depend on the hypothesis that we are able to properly classify all the sources to obtain a pure sample of galaxies. If this is not the case, the resulting contamination would add an extra level of uncertainty in our results that is not captured by our tests.

5.1. Point estimates

The results on the full spectroscopic sample presented in Fig. 4 show that not all template-fitting methods provide similar results. Although Le Phare, CPz, and Phosphoros implement almost exactly the same algorithm, Le Phare's results differ from these of the other codes, having slightly better results, with Fig. 3 showing a strong similarity between Phosphoros and CPz outputs. The differences are therefore due to details of the configuration of the methods (i.e., data-driven, instead of code-driven differences) and we have checked that Phosphoros can reproduce almost exactly the Le Phare results if run under identical conditions. The first difference are the templates, Le Phare is including two additional templates (generated with an exponentially declining SFH) in addition to the 31 COSMOS template of Ilbert et al. (2009) that Phophoros and CPz use. A second difference is the point-estimate definition. Both CPz and Phosphoros use the PDZ mode, but Le Phare uses the PDZ median. Also,

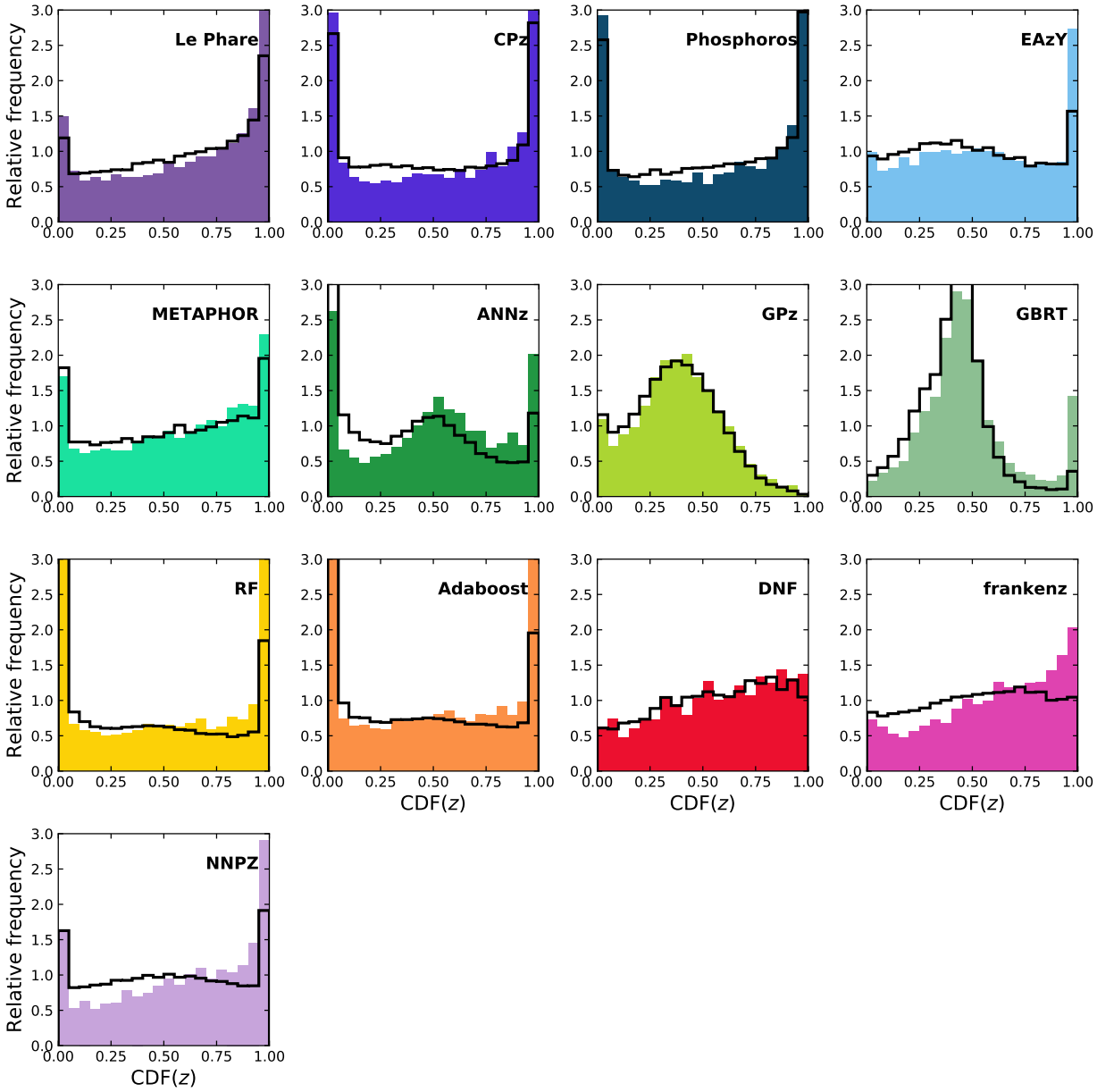

Fig. 7. Probability integral transform (PIT) plots for all the methods, for the USE = 1 population and the *Euclid* selection. Color histograms are the results for the weighted spectroscopic sample, while the black lines are the histograms for the L15 sample.

we note that `LePhare` applied an absolute magnitude cut when running, added systematic errors to the magnitude errors, and applied a rejection for sources with overly broad PDZs.

For `CPz` and `Phosphoros`, having made roughly the same configuration choices, we see that they yield very similar results, as can be seen in Figs. 3 and 4. The main difference between `LePhare` or `CPz` and `Phosphoros` is the point estimate definition. This mostly impacts the point estimate metrics, but is less relevant in the rest of the analysis using PDZs (see Sect. 5.2). Finally, the differences between `LePhare` and `Phosphoros` show that there is some room for improvement in the configuration of the algorithms.

`EAzY` is a bit distinct from the other template-fitting codes: it uses a different set of templates than the 31 COSMOS templates, which it combines to fit the data; it uses a prior on the magnitudes in the *r*-band, which is not set by the other template-fitting codes; and it applies a different rejection scheme than the other codes, based on the odds of a PDZ being single-peaked. The results presented in Fig. 3 are different than the results of the other template-fitting codes for these reasons. Despite these differences, it has similar performance to `LePhare`, as seen in Fig. 4.

We see in the top left panel of Fig. 4 that, for the whole sample, the lowest σ and η values are achieved by machine-learning algorithms (specifically `Adaboost` and `aNNz`). The rejection of the less reliable estimates can greatly improve the results for the point estimates. For example, `METAPHOR` sees its outlier fraction drop by about a factor of 6, and its scatter reduced by 25% when applying a rejection based on its USE flag. Most rejection schemes seem to efficiently identify outliers, but have only a small effect on the scatter. However, the improvement in the outlier fraction comes at a price for completeness, since the most precise method after the rejection of flagged objects (`METAPHOR`) discards 1/3 of all the sources and the second one (`DNF`) rejects half of them. Figure 4 shows also that the *Euclid* selection leads to an improvement of the outlier fraction, but mainly for the

Table 3. Median continuous ranked probability score (CRPS) for the different algorithms using the *Euclid* selection for the weighted spectroscopic and L15 samples.

	Spec. sample	L15 sample
LePhare	0.057	0.056
CPz	0.087	0.091
Phosphoros	0.082	0.083
EAzY	0.050	0.048
METAPHOR	0.036	0.034
ANNz	0.115	0.124
GPz	0.116	0.113
GBRT	0.166	0.165
RF	0.107	0.119
Adaboost	0.078	0.090
DNF	0.076	0.072
frankenz	0.071	0.072
NNPZ	0.084	0.081
Bias (0.06−0.18)	0.040–0.160	–
Scatter (0.15−0.7)	0.036–0.164	–

Notes. The last two rows present the CRPSs provided for sources in two cases: with infinitely precise Dirac PDZs but with a bias in range 0.06−0.18; and with absolutely accurate (bias = 0) Gaussian PDZs with a scatter in range 0.15−0.7.

methods that do not make any rejection of possibly wrong results on their own. This indicates that a large fraction of the outliers are located in redshift ranges outside of the *Euclid* cosmic-shear target region.

When weighting the results using the Lima et al. (2008) weighting scheme or using the L15 sample, we see that most of the machine-learning results degrade greatly if the methods do not apply strong rejection. Training is indeed very poor in areas of color space that have a large weight. This can be mitigated if algorithms are able to identify and reject objects in these areas, which is especially the case for METAPHOR. On the other hand, Adaboost, RF, and ANNz are strongly penalized by the absence of rejection in their configurations. Another mitigation measure is the use of L15 photo-zs in the training, as is the case for GBRT and NNPZ. For these examples, the results on the L15 sample are even better than those on the color-space weighted sample.

Color-space weight and L15 scores are mostly similar, except for Adaboost, RF, and ANNz, which shows that these methods are able to make reasonable predictions even with few training objects, but that there are significant areas of the color space without any spec-zs. This could mean that the color-space weights are overestimating the results that the methods would have on the full photometric sample. An alternative explanation is that this could be due to bad L15 photo-zs, since the comparison between L15 photo-zs and spec-zs in the validation catalog shows a scatter $\sigma = 0.013$ and an outlier fraction $\eta = 11.0\%$, this explanation cannot be excluded. However the consistency of the L15 and color-space results for most algorithms show that these errors, if they are significant, happen essentially where spec-z coverage is scarce. Template-fitting results do not seem to suffer as strongly as machine-learning results when applied to the color-space weighted or the L15 samples. Template-fitting appears to be able to provide sensible results even in the areas of the color space not covered by spectroscopic-redshifts, but we point out that template-fitting methods might perform well in these areas of color space because L15 photo-zs used the same algorithm and the same templates as LePhare, Phosphoros, and CPz. However, differences in depth and wavelength coverage somewhat mitigate this issue.

5.2. PDZs

Although PIT plots and CRPSs have been used in recent works as PDZ quality indicators (e.g., Tanaka et al. 2018; Pasquet et al. 2019), they are not very useful indicators of the precision of the PDZs. CRPS is sensitive to both bias and scatter, in a way that makes the two effects difficult to disentangle. PIT checks that the individual spec-zs can be on average drawn from the PDZs, but it does not say anything about the quality of the predictions. Schmidt et al. (2020) give the example of a method without any predictive power, but with a perfect PIT, meaning that a method providing a perfect PIT plot can lead to a bad FoM. The same behavior can be expected from the CRPS, since the same CRPS values dominated either by the bias or the precision of the PDZs will provide different FoMs. Nevertheless the general shapes of the PIT plots give some indications of whether the PDZs are over- or under-dispersed, biased, or outliers. Most PIT plots in Fig. 7 appear reasonable, with the exception of GPz, GBRT, and ANNz. One the other hand, none of the PIT histograms are flat. Some methods still manage to provide fairly flat but biased PIT plots, especially for the L15 sample (like frankenz), or are slightly concave in the center and with small peaks toward the edges (like NNPZ or EAzY). This means that no method can produce PDZs that are in total agreement with the spec-z or the L15 photo-z distributions. This is not a fatal issue, however, since there are ways to correct the PDZs in order to flatten the PIT plot (Bordoloi et al. 2010; Gomes et al. 2018) and to correct for most of the bias. In addition, the *Euclid* science goals do not require the determination of the true $n(z)$, but only of the average redshifts in the tomographic bins, which is a significantly less ambitious goal that can be reached, for example, using self-organizing maps as proposed by Masters et al. (2015).

In the context of the *Euclid* mission, we define new estimators of the photo-z precision that are insensitive to the bias. The *Euclid* requirements are expressed using the F_{005} and F_{015} definitions, which consider the PDZs around the mode of the distributions, making these metrics sensitive only to the precision (the width) of the PDZs. They can also be easily associated with the scatter and outlier fraction of point estimates. For these reasons, we focus on the F_{005} and F_{015} measurements for the different methods. The binning of the results in tomographic bins presented in Figs. 5 and 6 allows us to see more clearly what was hinted in Fig. 3, namely that strongly-rejective methods are mainly rejecting sources with redshifts $z > 1$. This is the case for METAPHOR, which does not provide any results above the bin at $z = 1.2-1.4$, neither for the weighted spectroscopic sample, nor the L15 sample. However, this strong rejectivity results in high scores for the metrics in the domain in which results are provided. Figures 5 and 6 show that the methods that have poor results in their PIT plots and CRPSs do not perform well on the F_{005} and F_{015} metrics (e.g., GBRT, GPz, or ANNz). Figure 5 shows that machine-learning methods tend to perform better than template-fitting ones in the redshift range of $z < 0.8$, but perform worse above this redshift. Using the L15 sample (Fig. 6), the gap between the results of machine-learning approaches and those of template-fitting is larger than that obtained from the spectroscopic sample. This indicates that (perhaps unsurprisingly) the machine-learning algorithms also have more difficulty in providing sensible PDZs for sources that are rarely or not at all represented in the training sample. An increase in the redshift coverage of the color space is needed to more properly train the machine-learning methods. Ongoing and future spectroscopic survey programs (e.g., C3R2, Masters et al. 2017, 2019; Euclid Collaboration 2020) will increase the color

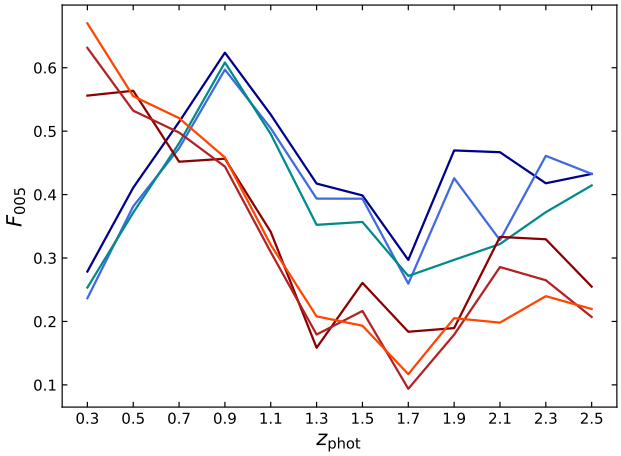

Fig. 8. F_{005} plot on the weighted spectroscopic sample showing the impact of the definition of the point estimates used to sort the sources into the redshift bins for `Phosphoros` and `Adaboost`.

space coverage with high-quality spectroscopic redshifts, thus the performance of machine-learning algorithms is expected to improve over time due to a better training sample. However, it is not clear that the number of spec-zs will be sufficient to both train the machine-learning methods and calibrate the bias of the photo-zs without introducing new sources of bias.

Template-fitting codes use an explicit model of the galaxy SEDs, and thus they provide better results at high redshift than machine-learning algorithms, which rely on training sources in this regime. However, at redshifts $z < 0.5$, all the template-fitting methods are outmatched by machine-learning methods. The superior results of machine-learning approaches at low redshifts show that the photometry does contain enough information to constrain the photo-zs. Nevertheless template-fitting methods have trouble in this region. This may result from a lack of valid templates at low redshift, or it may be due to a lack of proper priors, which are present in machine-learning methods in an implicit way due to the training data set containing mostly sources with low redshifts.

In Sect. 5.1, we explained that different definitions of point estimates can lead to differences in results. Our PDZ metrics are still sensitive to the point estimate variations, since we use them to sort the sources within the tomographic bins. Figure 8 shows an example of the impact of the definition of the point estimates on the F_{005} fraction for `Phosphoros` and `Adaboost`. We observe some differences between the results, mostly at redshifts $z > 1$. In that range of redshift, the mode seems to be the point-estimate that provides the best results. For the $z < 1$ redshift range, we see very little variation of the results with the definition of the point estimates.

5.3. Maximizing the proper metric

In the context of *Euclid*, the metric that is maximized is the dark-energy figure of merit (see Laureijs et al. 2011 for a detailed description). The dark energy FoM increases with the quality of the weak-lensing signal, and this signal depends on the quality of the photo-zs, but also on the number of sources for which the photo-zs are measured[5]. The requirement presented in the *Euclid* Red Book is that the galaxy density must be over 30 galaxies per arcmin^2.

[5] It clearly also depends on other parameters, but we focus here on the effects on which the photo-z algorithms have influence.

The results presented in Sects. 4.1 and 4.2 show that a rejection of the sources on which to carry out the analysis allows the methods to improve the precision of the redshifts. However, the F_{005} and F_{015} metrics are not sensitive to the loss of information resulting from this rejection. The same problem is true for PIT and CRPS. Figures 5 and 6 show that some methods, such as `METAPHOR`, leave some tomographic bins completely unpopulated. This means that no weak-lensing analysis can be performed at these redshifts, resulting in a strong loss of FoM and a failure to meet the *Euclid* mission requirements if such drastic rejection is made.

We use two methods of averaging the F_{005} and F_{015} metrics over the tomographic bins (Fig. 9). First, the weight applied to F_{005} and F_{015} in each bin is the number of objects put in this bin by a given photo-z method, that is,

$$\langle F_{0XX} \rangle = \frac{1}{N_{\mathrm{USE=1}}} \sum_{i}^{\mathrm{bins}} F_{0XX,i} N_{\mathrm{sources},i}, \tag{6}$$

where $F_{0XX,i}$ is either F_{005} or F_{015} (or any other desired value) in a bin i, $N_{\mathrm{sources},i}$ is the number of sources in a bin and $N_{\mathrm{USE=1}}$ is the total number of sources in all the bins. These weights roughly reproduce the standard estimators for point estimates σ and η in the case of PDZs, since they are averaged over all objects. The $\langle F_{0XX} \rangle$ metric does not penalize methods with strong rejection because the empty bins have null weight in the average computation, thus $\langle F_{0XX} \rangle$ does not reflect the negative impact that underpopulated, or even empty, tomographic bins can have on the weak-lensing analysis. Using this average, the best methods seem to be `METAPHOR`, `Le Phare`, and `Phosphoros`.

Another way to produce an average $\langle F_{0XX} \rangle$ would be to assume that each tomographic bin has the same weight in the weak-lensing signal, which translates into unweighted averages of F_{005} and F_{015}. This would give a penalty to methods rejecting all objects in a given bin or to methods that are particularly poor in some redshift range (typically machine-learning at high z). However, it would does not impact results with underpopulated bins that could obtain good F_{005} and F_{015} values, but not enough sources to improve the weak-lensing analysis results. For this reason, the metric must take into account the population of the tomographic bins. To do so, a correction is introduced that depends on the fraction of objects correctly assigned to the bin:

$$\langle F_{0XX} \rangle_{\mathrm{corrected}} = \frac{1}{N_{\mathrm{bins}}} \sum_{i}^{\mathrm{bins}} F_{0XX,i} \sqrt{\frac{N_{\mathrm{good},i}}{N_{\mathrm{true},i}}}, \tag{7}$$

where $N_{\mathrm{good},i}$ is the number of sources that have been correctly placed in the bins i, and $N_{\mathrm{true},i}$ is the true number of sources in bin i (see Fig. E.1 for the values of the fractions per bin). The square root is applied to reproduce the dependency of the increase in precision with the number of objects. Using the fraction of "good" sources compared to the number of "true" sources in the bin penalizes underpopulated but not empty bins with high fraction values. It also ignores outliers falling in the bins, which could artificially boost the scores of the bins. Figure 9 (right panel) shows the result of this correction. Template-fitting methods (`Le Phare` and `Phosphoros`) present the best results, but some machine-learning methods being less penalized, such as `Adaboost` and `NNPZ`, also yields good performance. Nevertheless, this correction is a simple and intuitive way to estimate the trade-off between the precision of the photo-zs and the number of sources considered, but the proper metrics to consider here would take into account the weight of each sources and tomographic bins in the estimation of the weak-lensing signal.

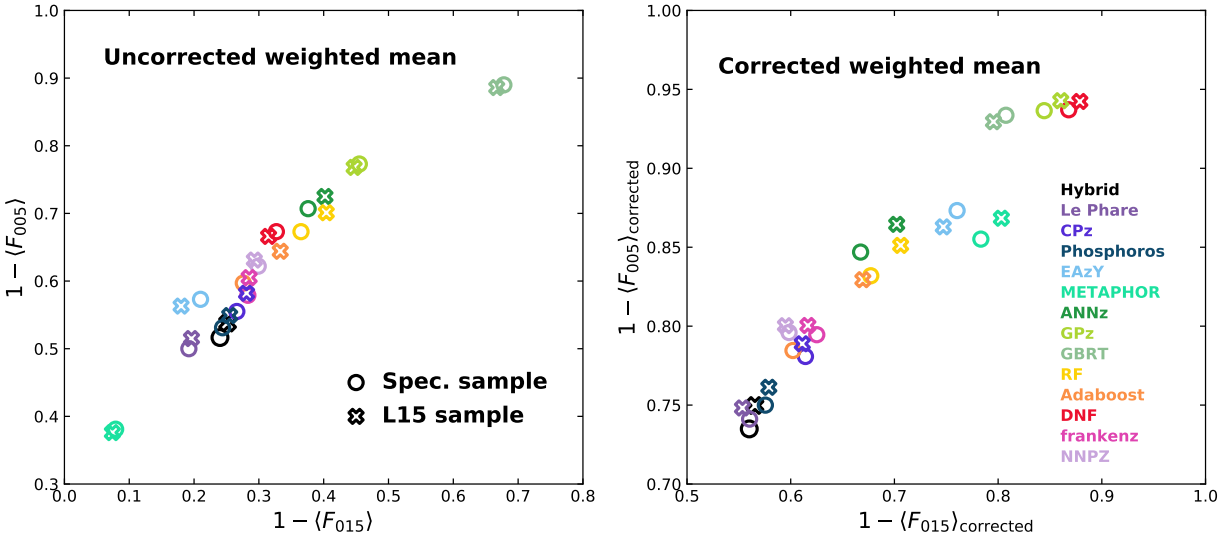

Fig. 9. PDZs metrics summarized by averaging the F_{005} and F_{015} values on all the bins with different weighting schemes. The axes are $1-\langle F_{005}\rangle$ and $1-\langle F_{015}\rangle$ to mimic the usual σ–η plots in Fig. 4. *Left*: results per bin weighted by the fraction of sources in the bin compared to the total number of sources kept by each methods (see Eq. (6)). *Right*: results of all the methods when correcting the $\langle F_{005}\rangle$ and $\langle F_{015}\rangle$ of each bin by the square root of the ratio of good sources in the bins to the number of sources that truly belong to the bin (see Eq. (7)). In each plot we include the results for the hybrid method (in black, see Sect. 5.4) for the weighted spectroscopic sample on the L15 sample.

It would be desirable to apply a penalty similar to that used in Eq. (7) for the PIT and CRPS metrics. Unfortunately, there is no sensible way to estimate how the loss of sources would affect them, and neither the CRPS, nor any statistics derived from the PIT can be unambiguously translated into a FoM.

5.4. Improving the results

Each methods has its advantages and disadvantages, and thus performs efficiently in different regimes. Machine-learning methods are based on a training sample and their results depend strongly on the quality of this training. Template-fitting methods do not have this problem and perform relatively well for sources in regions of the color space with a sparse redshift coverage. However, Fig. 5 shows that they can be outmatched by machine-learning in the well covered regions of the color space. As mentioned earlier, both types of method can be improved separately (see Sect. 5.2). However, Fig. 5 also shows that some methods (such as METAPHOR) are able to substantially improve the precision of their results by accurately predicting when a result is a probable outlier. In Sect. 5.3 we see that non-rejective template-fitting methods (such as Phosphoros or Le Phare) are performing well with a metric approximating the effect of photo-zs on the weak-lensing analysis, whereas very precise but highly rejecting methods (such as Metaphor or DNF) not providing results above redshift 2 are incompatible with the goals of the weak-lensing analysis. Nevertheless, the ability of METAPHOR to predict outliers could be used to improve the results. For that, we first must check if METAPHOR really surpasses other methods for the objects for which it provides results. Figure 10 presents the F_{005} curve for Phosphoros restricted to sources that have a USE flag 1 with METAPHOR. We see that the Phosphoros results are greatly improved up to the point where the rejection discards all the sources, above $z = 1.5$. Also, we note that the results of METAPHOR are only better than the results of Phosphoros with the same rejection in the first two bins of the *Euclid* redshift range. Thus, to improve results, we propose a hybrid photo-z algorithm as follow: we use the results of METAPHOR when its USE flag is 1 and its predicted photo-z is below $z = 0.6$, otherwise we use the Phosphoros results if not.

The results of this hybrid method are presented in Fig. 10 with a F_{005} plot and a scatter plots. The scatter plot in Fig. 10 shows that the hybrid approach has similar scatter, but a slightly smaller outlier fraction than Phosphoros (see Table B.3), which results in an increase in the number of sources in the Euclid sample at low redshift. The results of the hybrid method are the best ones of all methods that do not reject any source. The resulting F_{005} curve in Fig. 10 differs from the Phosphoros one only in the two first bins, since above $z = 0.6$ only Phosphoros results are used. The curve still remain under the METAPHOR one, but a solution is provided for all the sources in the sample. We can see the improvement it brings in Fig. 9, that shows the results of this approach in the averaged F_{0XX} metrics. The corrected mean F_{005} of this hybridization is $\langle F_{005}\rangle_{\rm corrected} = 0.27$ on the weighted spectroscopic sample. This result is better than all the results presented in Fig. 9. The increase compared to Phosphoros is only due to the improvement in F_{005} in the first two bins.

This method of combining the results of machine-learning and template-fitting seems promising (e.g., Brodwin et al. 2006; Duncan et al. 2018) and should be explored further, possibly with a better criterion for selecting the predictions from the machine-learning or the template-fitting algorithms.

6. Summary and conclusions

Thirteen different photo-z methods, both template-fitting and machine-learning based, have been tested on *Euclid*-like data. Each method has provided each source a point estimate redshift, a PDZ, and a USE flag, allowing them to reject sources considered problematic. Their results have been compared through different metrics with the aim of assessing the impact of the provided photo-zs on the *Euclid* cosmic-shear analysis. For this reason, we analyze the results for galaxies in the 0.2–2.6 photo-z range only. The tests we have conducted here have therefore little relevance for the study of high-redshift galaxies, for instance. We have further assumed that a proper classification between stars, galaxies and AGNs has already been done, and that the photo-zs can be calibrated independently of the photo-z algorithm.

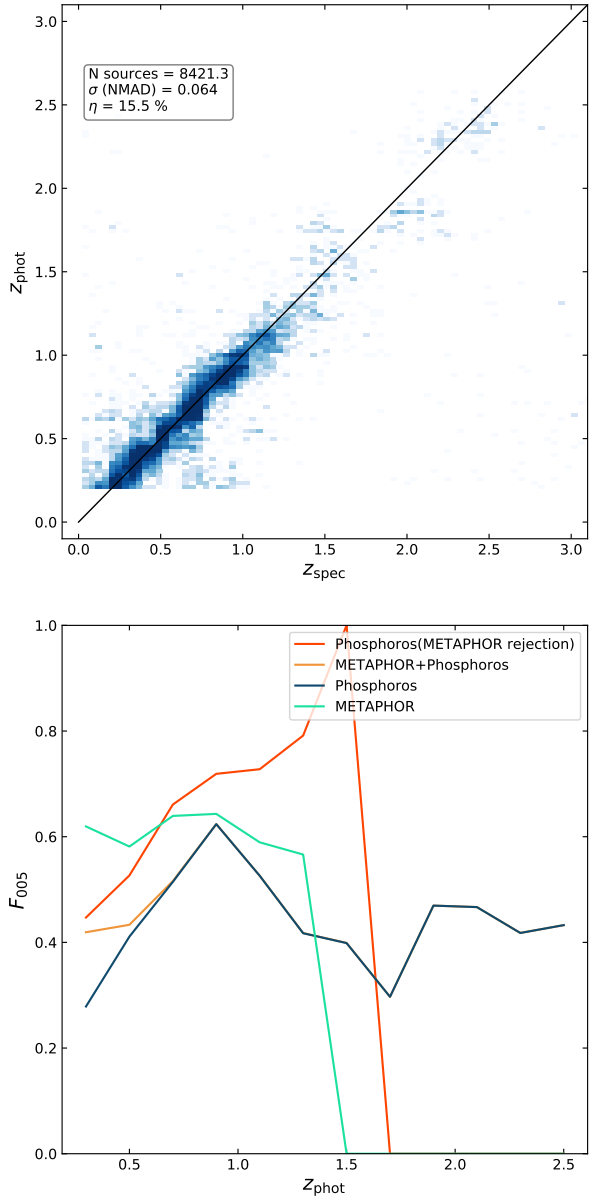

Fig. 10. *Top*: scatter plot of results of the combination of METAPHOR and Phosphoros point estimates on the weighed spectroscopic sample and the *Euclid* selection. *Bottom*: F_{005} for METAPHOR and Phosphoros results along with the combination of their results. The red curve is the results of Phosphoros when the METAPHOR rejection scheme is applied to it. The orange curve shows the combination of the results of METAPHOR and Phosphoros in the first two bins, then the curves merges with the typical Phosphoros one.

The results show that adopting stringent rejection criteria can be very efficient in reducing the outlier fraction. Some methods are quite successful in accurately identifying sources with reliable photo-zs. However, the drawback of such rejection is a loss of completeness for further analysis, which can be incompatible with some science goals, in particular weak-lensing tomography.

To assess the quality of PDZs, the PIT plot and CRPS are standard metrics. They must be considered together, since PIT only indicates whether the true redshifts are collectively compatible with being drawn from the PDZs, and CRPS is sensitive to both the bias and scatter of the results. However, its sensitivity to the bias, which cannot be disentangled from the effect of the scatter, is not suited for our analysis that only focuses on the precision of the results. This leads us to define the fraction metrics (F_{005} and F_{015}) related to the *Euclid* requirements on the precision of the PDZs. The fraction metrics can also be corrected to take in account the loss in completeness that is due to rejection schemes of the different methods.

Analysis of the PDZ results shows that producing sensible PDZs is not straightforward for machine-learning methods, as several of them do not manage to provide good PDZs, regardless of the indicator used to assess their quality. Machine-learning methods also struggle to make good predictions over large areas of the color space, in particular for $z > 1$ or regions scarcely covered by spectroscopic information, even though the COSMOS training sample is one of the most complete spec-z samples currently available.

However, in regions of color space well covered by spec-zs, machine-learning methods (e.g., METAPHOR or Adaboost) seem to perform the best. With an appropriate spec-z sample, they could outmatch all the other methods. However, the construction of a perfect training sample covering the full color space at the limiting depth of the surveys with sufficiently numerous spec-z, remains intractable. Using L15 photo-zs for this purpose is a possible compromise, as shown by NNPZ in particular.

Template-fitting methods show more consistent results than machine-learning over the full photometric sample; however, they seem unable to use the full information contained in the photometry at low redshifts. The reason for this behavior must be understood whether it is a lack of templates or a better definition of priors in order to improve these methods.

Taking into account the properties of the output photo-zs, the driver of the choice of algorithm is the use made of them. The metrics used to compare the results of the algorithms depend on the purpose of the photo-zs and must reflect the impacts they will have on the science case foreseen. For weak-lensing studies, completeness is needed, and template-fitting appears to perform best when assessing both the precision of the photo-zs and their numbers. However, if high precision and purity are required, then machine-learning seems better in those aspects, especially when they implement rejection of poor predictions.

Thanks to the capability of rejecting probable outliers, we can overcome the limits of both approaches and combine the high precision of machine-learning and the completeness of template-fitting. This combination of results shows better average photo-z precision than any method alone, while preserving the completeness of the considered sample of galaxies, hence solving the issue of the loss of sources, which impacts negatively the weak-lensing analysis and the *Euclid* dark energy FoM. Further work is required to determine the optimal combination between template-fitting and machine-learning algorithms.

Acknowledgements. GD thanks Douglas Scott for his very helpful comments on the manuscript. GD and AG acknowledge the support from the Sinergia program of the Swiss National Science Foundation. Part of this work was supported by the German *Deutsche Forschungsgemeinschaft, DFG* project number Ts 17/2–1. MB acknowledges the financial contribution from the agreement *ASI/INAF 2018-23-HH.0, Euclid ESA mission – Phase D* and the *INAF PRIN-SKA 2017 program 1.05.01.88.04*. SC acknowledges the financial contribution from FFABR 2017. 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 Ciência 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

Abdalla, F. B., Banerji, M., Lahav, O., & Rashkov, V. 2011, MNRAS, 417, 1891
Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, PASJ, 70, S4
Akeson, R., Armus, L., Bachelet, E., et al. 2019, ArXiv e-prints [arXiv:1902.05569]
Almosallam, I. A., Lindsay, S. N., Jarvis, M. J., & Roberts, S. J. 2016a, MNRAS, 455, 2387
Almosallam, I. A., Jarvis, M. J., & Roberts, S. J. 2016b, MNRAS, 462, 726
Amaro, V., Cavuoti, S., Brescia, M., et al. 2019, MNRAS, 482, 3116
Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, MNRAS, 310, 540
Arnouts, S., Moscardini, L., Vanzella, E., et al. 2002, MNRAS, 329, 355
Baum, W. A. 1962, in Problems of Extra-Galactic Research, ed. G. C. McVittie, IAU Symp., 15, 390
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bohlin, R. C. 2016, AJ, 152, 60
Bolzonella, M., Miralles, J. M., & Pelló, R. 2000, A&A, 363, 476
Bordoloi, R., Lilly, S. J., & Amara, A. 2010, MNRAS, 406, 881
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Breiman, L. 2001, Mach. Learn., 45, 5
Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, ApJ, 772
Brodwin, M., Brown, M. J. I., Ashby, M. L. N., et al. 2006, ApJ, 651, 791
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Cavuoti, S., Brescia, M., De Stefano, V., & Longo, G. 2015, Exp. Astron., 39, 45
Cavuoti, S., Amaro, V., Brescia, M., et al. 2017, MNRAS, 465, 1959
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
Connolly, A. J., Csabai, I., Szalay, A. S., et al. 1995, AJ, 110, 2655
Dahlen, T., Mobasher, B., Faber, S. M., et al. 2013, ApJ, 775, 93
Dawid, A. P. 1984, J. R. Stat. Soc. Ser. A, 147, 278
de Jong, J. T. A., Kuijken, K., Applegate, D., et al. 2013, The Messenger, 154, 44
De Vicente, J., Sánchez, E., & Sevilla-Noarbe, I. 2016, MNRAS, 459, 3078
Desai, S., Armstrong, R., Mohr, J. J., et al. 2012, ApJ, 757, 83
Desai, S., Mohr, J. J., Henderson, R., et al. 2015, J. Instrum., 10, C06014
Desai, S., Mohr, J. J., Bertin, E., Kümmel, M., & Wetzstein, M. 2016, Astron. Comput., 16, 67
D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111
Duncan, K. J., Jarvis, M. J., Brown, M. J. I., & Röttgering, H. J. A. 2018, MNRAS, 477, 5177
Euclid Collaboration (Guglielmo, V. et al.) 2020, A&A, 642, A192
Firth, A. E., Lahav, O., & Somerville, R. S. 2003, MNRAS, 339, 1195
Flaugher, B. 2005, Int. J. Mod. Phys. A, 20, 3121
Fotopoulou, S., & Paltani, S. 2018, A&A, 619, A14
Freund, Y., & Schapire, R. E. 1997, J. Comput. Syst. Sci., 55, 119
Friedman, J. H. 2001, Ann. Stat., 29, 1189
Galametz, A., Saglia, R., Paltani, S., Apostolakos, N., & Dubath, P. 2017, A&A, 598, A20
Giavalisco, M., Ferguson, H. C., Koekemoer, A. M., et al. 2004, ApJ, 600, L93
Gomes, Z., Jarvis, M. J., Almosallam, I. A., & Roberts, S. J. 2018, MNRAS, 475, 331
Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35
Hennig, C., Mohr, J. J., Zenteno, A., et al. 2017, MNRAS, 467, 4015
Hersbach, H. 2000, Weather Forecast., 15, 559
Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, A&A, 523, A31
Hogg, D. W., Cohen, J. G., Blandford, R., et al. 1998, AJ, 115, 1418
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
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Koekemoer, A. M., Aussel, H., Calzetti, D., et al. 2007, ApJS, 172, 196
Koo, D. C. 1985, AJ, 90, 418
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Lanzetta, K. M., Yahil, A., & Fernández-Soto, A. 1996, Nature, 381, 759
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints [arXiv:1110.3193]
Leauthaud, A., Massey, R., Kneib, J.-P., et al. 2007, ApJS, 172, 219
Lima, M., Cunha, C. E., Oyaizu, H., et al. 2008, MNRAS, 390, 118
Loh, E. D., & Spillar, E. J. 1986, ApJ, 303, 154
Lupton, R. H., Gunn, J. E., & Szalay, A. S. 1999, AJ, 118, 1406
Marchesi, S., Civano, F., Elvis, M., et al. 2016, ApJ, 817, 34
Massey, R., Stoughton, C., Leauthaud, A., et al. 2010, MNRAS, 401, 371
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
McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, A&A, 544, A156
Mohr, J. J., Armstrong, R., Bertin, E., et al. 2012, in Software and Cyberinfrastructure for Astronomy II, Proc. SPIE, 8451, 84510D
Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Polletta, M., Tajer, M., Maraschi, L., et al. 2007, ApJ, 663, 81
Prevot, M. L., Lequeux, J., Maurice, E., Prevot, L., & Rocca-Volmerange, B. 1984, A&A, 132, 389
Puschell, J., Owen, F., & Laing, R. 1982, in Extragalactic Radio Sources, eds. D. S. Heeschen, & C. M. Wade, IAU Symp., 97, 423
Rau, M. M., Seitz, S., Brimioulle, F., et al. 2015, MNRAS, 452, 3710
Salvato, M., Ilbert, O., Hasinger, G., et al. 2011, ApJ, 742, 61
Salvato, M., Ilbert, O., & Hoyle, B. 2019, Nat. Astron., 3, 212
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020, MNRAS, 499, 1587
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Singal, J., Shmakova, M., Gerke, B., Griffith, R. L., & Lotz, J. 2011, PASP, 123, 615
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Soo, J. Y. H., Moraes, B., Joachimi, B., et al. 2018, MNRAS, 475, 3613
Tagliaferri, R., Longo, G., Andreon, S., et al. 2003, Neural Networks for Photometric Redshifts Evaluation (Berlin, Heidelberg: Springer), 2859, 226
Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, PASJ, 70, S9
Thomas, D., Steele, O., Maraston, C., et al. 2013, MNRAS, 431, 1383
Way, M. J., Foster, L. V., Gazis, P. R., & Srivastava, A. N. 2009, ApJ, 706, 623

[1] Department of Astronomy, University of Geneva, Ch. d'Écogia 16, 1290 Versoix, Switzerland
[2] Saudi Information Technology Company, Riyadh 12382, Saudi Arabia
[3] King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia
[4] Information Engineering, Parks Road, Oxford OX1 3PJ, UK
[5] School of Physics and Astronomy, Sun Yat-sen University, Guangzhou 519082, Zhuhai Campus, PR China
[6] INAF-Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
[7] Department of Physics and Astronomy, University of Missouri, 5110 Rockhill Road, Kansas City, MO 64110, USA
[8] Department of Physics "E. Pancini", University Federico II, Via Cinthia 6, 80126 Napoli, Italy
[9] INFN Section of Naples, Via Cinthia 6, 80126 Napoli, Italy
[10] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avenida Complutense 40, 28040 Madrid, Spain
[11] School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK
[12] Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK
[13] Aix-Marseille Univ., CNRS, CNES, LAM, Marseille, France
[14] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[15] Center for Astrophysics | Harvard & Smithsonian, 60 Garden St., Cambridge, MA 02138, USA
[16] INAF-Osservatorio Astronomico di Roma, Via Frascati 33, 00078 Monteporzio Catone, Italy
[17] Universitäts-Sternwarte München, Fakultät für Physik, Ludwig-Maximilians-Universität München, Scheinerstraße 1, 81679 München, Germany
[18] Sorbonne Universités, UPMC Univ. Paris 6 et CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 Bis Bd Arago, 75014 Paris, France
[19] Max Planck Institute for Extraterrestrial Physics, Giessenbachstr. 1, 85748 Garching, Germany
[20] INAF-Osservatorio Astronomico di Brera, Via Brera 28, 20122 Milano, Italy
[21] INAF-Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Piero Gobetti 93/3, 40129 Bologna, Italy

[22] SISSA, International School for Advanced Studies, Via Bonomea 265, 34136 Trieste, TS, Italy
[23] INFN, Sezione di Trieste, Via Valerio 2, 34127 Trieste TS, Italy
[24] INAF-Osservatorio Astronomico di Trieste, Via G. B. Tiepolo 11, 34131 Trieste, Italy
[25] Universidad de la Laguna, 38206 San Cristóbal de La Laguna, Tenerife, Spain
[26] Instituto de Astrofísica de Canarias, Calle Vía Làctea s/n, 38204 San Cristóbal de la Laguna, Tenerife, Spain
[27] Dipartimento di Fisica e Astronomia, Universitá di Bologna, Via Gobetti 93/2, 40129 Bologna, Italy
[28] INFN-Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[29] IFPU, Institute for Fundamental Physics of the Universe, Via Beirut 2, 34151 Trieste, Italy
[30] INAF-Osservatorio Astrofisico di Torino, Via Osservatorio 20, 10025 Pino Torinese, TO, Italy
[31] INFN-Sezione di Roma Tre, Via della Vasca Navale 84, 00146 Roma, Italy
[32] Department of Mathematics and Physics, Roma Tre University, Via della Vasca Navale 84, 00146 Rome, Italy
[33] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, 4150-762 Porto, Portugal
[34] Dipartimento di Fisica e Scienze della Terra, Universitá degli Studi di Ferrara, Via Giuseppe Saragat 1, 44122 Ferrara, Italy
[35] INAF, Istituto di Radioastronomia, Via Piero Gobetti 101, 40129 Bologna, Italy
[36] Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, 31400 Toulouse, France
[37] INFN-Sezione di Torino, Via P. Giuria 1, 10125 Torino, Italy
[38] Dipartimento di Fisica, Universitá degli Studi di Torino, Via P. Giuria 1, 10125 Torino, Italy
[39] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, 06304 Nice Cedex 4, France
[40] INAF-IASF Milano, Via Alfonso Corti 12, 20133 Milano, Italy
[41] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra, Barcelona, Spain
[42] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Tapada da Ajuda, 1349-018 Lisboa, Portugal
[43] Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain
[44] Institut d'Estudis Espacials de Catalunya (IEEC), 08034 Barcelona, Spain
[45] AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, 91191 Gif-sur-Yvette, France
[46] Observatoire de Sauverny, Ecole Polytechnique Fédérale de Lausanne, 1290 Versoix, Switzerland
[47] INAF-Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125 Firenze, Italy
[48] Centre National d'Etudes Spatiales, Toulouse, France
[49] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[50] University of Nottingham, University Park, Nottingham NG7 2RD, UK
[51] European Space Agency/ESRIN, Largo Galileo Galilei 1, 00044 Frascati Roma, Italy
[52] ESAC/ESA, Camino Bajo del Castillo, s/n, Urb. Villafranca del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain
[53] Univ. Lyon, Univ. Claude Bernard Lyon 1, CNRS/IN2P3, IP2I Lyon, UMR 5822, 69622 Villeurbanne, France
[54] University of Lyon, UCB Lyon 1, CNRS/IN2P3, IUF, IP2I, Lyon, France
[55] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande 1749-016, Lisboa, Portugal
[56] Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[57] Université Paris-Saclay, CNRS, Institut d'Astrophysique Spatiale, 91405 Orsay, France
[58] Department of Physics & Astronomy, University of Sussex, Brighton BN1 9QH, UK
[59] Astrophysics Group, Blackett Laboratory, Imperial College London, London SW7 2AZ, UK
[60] INFN-Bologna, Via Irnerio 46, 40126 Bologna, Italy
[61] Institut de Physique Nucléaire de Lyon, 4, Rue Enrico Fermi, 69622 Villeurbanne Cedex, France
[62] Aix-Marseille Univ., CNRS/IN2P3, CPPM, Marseille, France
[63] Department of Physics, University of Helsinki, PO Box 64, 00014 Helsinki, Finland
[64] Department of Physics and Helsinki Institute of Physics, University of Helsinki, Gustaf Hällströmin Katu 2, 00014 Helsinki, Finland
[65] Dipartimento di Fisica "Aldo Pontremoli", Universitá degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy
[66] INFN-Sezione di Milano, Via Celoria 16, 20133 Milano, Italy
[67] Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK
[68] Institute of Theoretical Astrophysics, University of Oslo, PO Box 1029, Blindern 0315, Oslo, Norway
[69] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[70] von Hoerner & Sulger GmbH, SchloßPlatz 8, 68723 Schwetzingen, Germany
[71] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany
[72] Institut d'Astrophysique de Paris, 98Bis Boulevard Arago, 75014 Paris, France
[73] Université de Genève, Département de Physique Théorique and Centre for Astroparticle Physics, 24 Quai Ernest-Ansermet, 1211 Genève 4, Switzerland
[74] NOVA Optical Infrared Instrumentation Group at ASTRON, Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands
[75] Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, 53121 Bonn, Germany
[76] Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham, DH1 3LE, UK
[77] Institut für Theoretische Physik, University of Heidelberg, Philosophenweg 16, 69120 Heidelberg, Germany
[78] Zentrum für Astronomie, Universität Heidelberg, Philosophenweg 12, 69120 Heidelberg, Germany
[79] Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, France
[80] INAF-IASF Bologna, Via Piero Gobetti 101, 40129 Bologna, Italy
[81] Université de Paris, 75013 Paris, France
[82] LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Université, 75014 Paris, France
[83] Space Science Data Center, Italian Space Agency, Via del Politecnico snc, 00133 Roma, Italy
[84] Institute of Space Science, Bucharest 077125, Romania
[85] Institute for Computational Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[86] INFN-Padova, Via Marzolo 8, 35131 Padova, Italy
[87] Dipartimento di Fisica e Astronomia "G.Galilei", Universitá di Padova, Via Marzolo 8, 35131 Padova, Italy
[88] Departamento de Física, FCFM, Universidad de Chile, Blanco Encalada 2008 Santiago, Chile
[89] Universidad Politécnica de Cartagena, Departamento de Electrónica y Tecnología de Computadoras, 30202 Cartagena, Spain
[90] Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA 91125, USA
[91] Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK
[92] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

Appendix A: The VIS simulation software

The VIS simulation software takes as input a high-resolution image (using the HST ACS $F814W$ in the COSMOS field in this specific case) and manipulates it in order to obtain a simulated image with the desired features (i.e., degrading the resolution to the expected VIS resolution and adding noise). The pipeline implements four processing steps executed in the following sequence:

- Mkkernel, generates an analytical (Gaussian) kernel according to the input image PSF and the PSF requested for the simulated one;
- Convolve, operates the convolution from the input image to the convolved one according to the previously generated kernel;
- Swarp, performs the rebinning of the convolved image to the required pixel scale;
- Mknoise, Gaussian noise is added in each pixel to reproduce the desired depth in the output.

The original ACS $F814W$ image has a non-uniform depth, and particular care has been devoted to the noise addition: Gaussian noise is added to each pixel according to a scaling factor that takes into account the pixel-to-pixel variation of the original image depth. The resulting rms map is an image with constant value in the portion covered by the observation and has a constant value of 10^{16} outside. The rms map value is the result of the following equation:

$$\mathrm{rms_{out}} = \frac{10^{0.4\,(\mathrm{ZP}-m_n)}}{S/N\ \sqrt{\pi}\ \frac{nFWHM}{2\,\mathrm{pxs}}}, \tag{A.1}$$

where m_n is the reference magnitude (at the given S/N) measured in n (1, 2, or 3) times the PSF FWHM. ZP and pxs are the zero-point and the pixelscale of the image, respectively. Where the original image has been found to be shallower than requested no Gaussian noise has been added and the rms value has not been modified. Currently, photon noise from the sources is not included.

Appendix B: Point estimate metric tables

Table B.1. Point estimate statistics for the spectroscopic sample.

	σ_{all}	η_{all} [%]	$N_{\mathrm{USE=1}}$	$\sigma_{\mathrm{USE=1}}$	$\eta_{\mathrm{USE=1}}$ [%]
Le Phare	0.046	12.0	11 377	0.043	8.1
CPz	0.066	15.5	10 841	0.066	15.6
Phosphoros	0.066	15.8	12 463	0.066	15.8
EAzY	0.058	15.4	9594	0.047	6.2
METAPHOR	0.051	16.6	8302	0.037	2.8
ANNz	0.048	10.0	12 463	0.048	10.0
GPz	0.078	14.2	10 676	0.069	9.3
GBRT	0.058	9.7	12 311	0.058	9.2
RF	0.052	11.5	12 463	0.052	11.5
Adaboost	0.046	9.2	12 463	0.046	9.2
DNF	0.055	12.2	5520	0.041	5.9
frankenz	0.068	28.3	9661	0.042	8.8
NNPZ	0.061	12.1	12 463	0.061	12.1

Notes. The scatter (σ) and the outlier fraction (η) are given in the case of no rejection with subscript "all" (i.e., 12 463 sources), and in the case of rejection with the USE flag with subscript "USE = 1". In the second case, the number of selected source is also displayed ($N_{\mathrm{USE=1}}$).

In Sect. 4.1, we present the results of the different methods in several conditions, using multiple selections (e.g., USE flag or *Euclid* selection) and comparing the photo-z's to different reference redshifts. These results are summarized in Fig. 4, using the values compiled in Tables B.1–B.4 in this section.

Table B.1 contains the scatter (σ_{all}) and outlier fraction (η_{all}) for all the methods, considering the complete spectroscopic shear sample (12463 sources) present in the validation catalog. This table also shows for each method the number of sources remaining after the USE flag selection is applied ($N_{\mathrm{USE=1}}$) and the $\sigma_{\mathrm{USE=1}}$ and $\eta_{\mathrm{USE=1}}$ associated with this selection.

The results for the *Euclid* selection (i.e., being part of shear sample, with photo-z in the range 0.2–2.6, and USE = 1) are listed in Table B.2. The column N_{Euclid} shows the number of sources for each method. The scatter σ_{Euclid} and the outlier fraction η_{Euclid} are also reported.

Table B.3 presents the results for the *Euclid* selection after re-weighting it to be more representative of the photometric sample, using the Lima et al. (2008) scheme. For each method we provide $N_{\mathrm{color-space}}$, which is the sum of the computed weights of all the selected sources, along with the weighted scatter $\sigma_{\mathrm{color-space}}$ and outlier fraction $\eta_{\mathrm{color-space}}$.

Table B.2. Point estimate statistics for the *Euclid* sample.

	N_{Euclid}	σ_{Euclid}	η_{Euclid} [%]
Le Phare	10 607	0.041	6.9
CPz	8985	0.055	9.7
Phosphoros	10 140	0.055	8.7
EAzY	9286	0.046	6.2
METAPHOR	7865	0.036	2.7
ANNz	12 012	0.048	10.1
GPz	10 208	0.068	9.2
GBRT	11 978	0.057	9.3
RF	11 955	0.05	10.5
Adaboost	12 008	0.045	9.0
DNF	5101	0.042	5.8
frankenz	8870	0.041	8.1
NNPZ	11 501	0.059	11.1

Notes. Number of sources N_{Euclid}, scatter σ_{Euclid}, and outlier fractions η_{Euclid} are provided.

Table B.3. Point estimate statistics for the color-space weighted-sample.

	$N_{\mathrm{color-space}}$	$\sigma_{\mathrm{color-space}}$	$\eta_{\mathrm{color-space}}$ [%]
Le Phare	7644.0	0.056	13.4
CPz	8569.7	0.077	21.1
Phosphoros	8084.1	0.067	17.1
EAzY	5772.5	0.062	13.2
METAPHOR	3039.6	0.04	3.1
ANNz	10564.7	0.091	26.1
GPz	5391.3	0.082	13.7
GBRT	10280.1	0.085	22.5
RF	10657.5	0.114	32.6
Adaboost	10021.2	0.075	20.9
DNF	2326.2	0.053	9.3
frankenz	7807.7	0.069	22.0
NNPZ	10112.7	0.082	22.4

Notes. The sum of the weights of sources in each method selection $N_{\mathrm{color-space}}$, scatter $\sigma_{\mathrm{color-space}}$, and outlier fractions $\eta_{\mathrm{color-space}}$ are provided.

Another sample being somewhat representative of the full photometric one is the shear sample with the 30-band photo-zs from Laigle et al. (2016). Using these redshifts as reference, the scatter $\sigma_{30\text{-bands}}$ and outlier fraction $\eta_{30\text{-bands}}$ are presented in Table B.4. The number of selected sources $N_{30\text{-bands}}$ for each method is also shown.

Table B.4. Point estimate statistics for the L15 sample.

	$N_{30\text{-bands}}$	$\sigma_{30\text{-bands}}$	$\eta_{30\text{-bands}}$ [%]
LePhare	36 842	0.055	11.7
CPz	43 258	0.08	20.0
Phosphoros	38 649	0.069	16.5
EAzY	25 114	0.059	9.8
METAPHOR	13 830	0.04	2.7
ANNz	52 094	0.114	32.3
GPz	23 207	0.082	13.7
GBRT	50 980	0.081	19.1
RF	52 391	0.136	37.4
Adaboost	49 718	0.096	26.9
DNF	9694	0.052	8.5
frankenz	42 808	0.07	19.1
NNPZ	51 047	0.081	19.9

Notes. The number of sources $N_{30\text{-bands}}$, scatter $\sigma_{30\text{-bands}}$, and outlier fractions $\eta_{30\text{-bands}}$ are provided.

Appendix C: photo-z versus 30-band photo-z

photo-zs provided by all the tested methods are compared to reference redshifts to examine the performance of the codes. In Sect. 4.1, we present a comparison between spec-zs and photo-zs, specifically shown in Fig. 3. Figure C.1 makes a comparison between the code photo-zs and 30-band photo-zs of Laigle et al. (2016), which better represent the full photometric sample. The resulting metrics associated with these plots are presented in Table B.4.

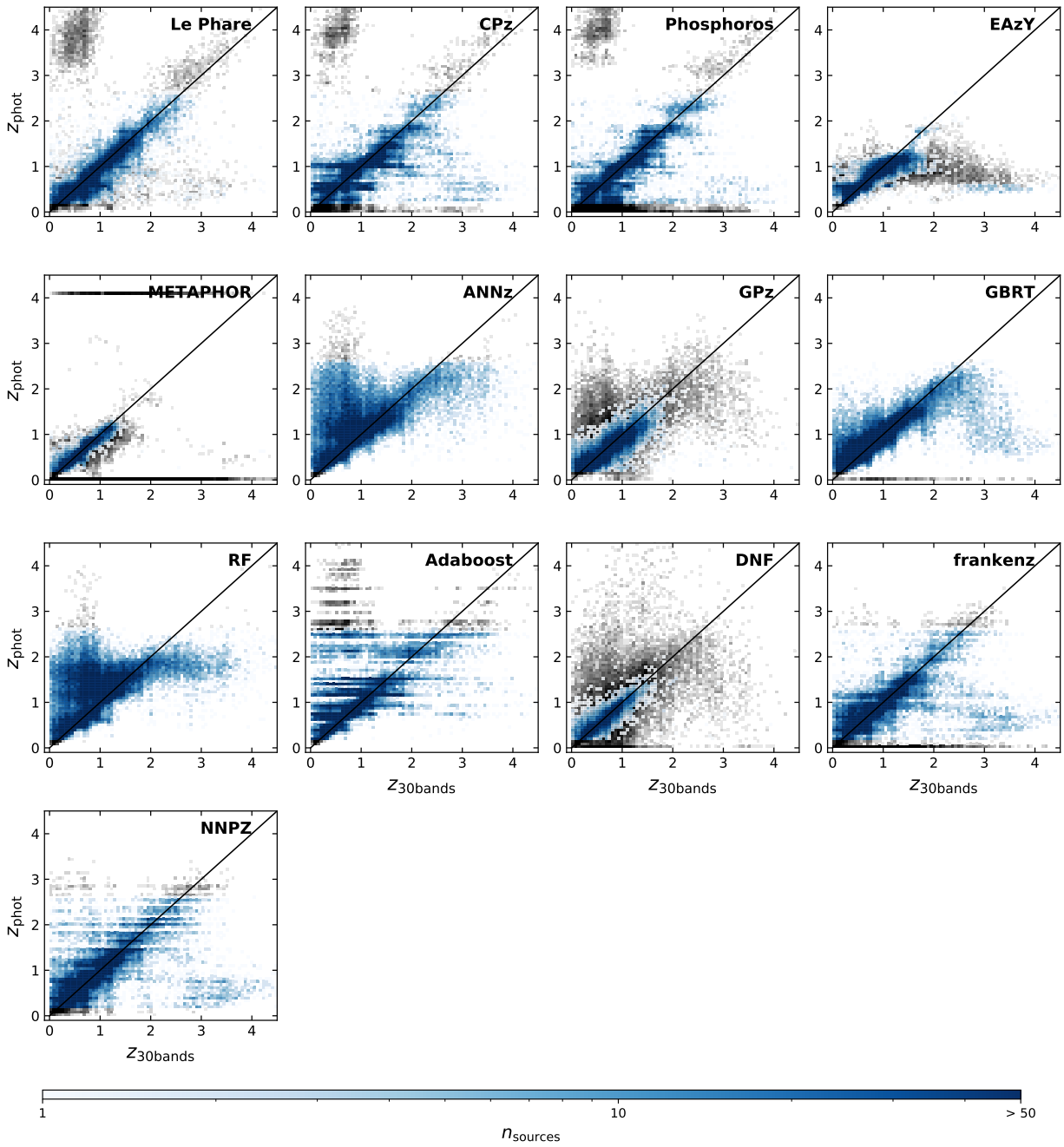

Fig. C.1. Photometric redshifts between $0.2 < z \leq 2.6$ measured with all the methods compared to the Laigle et al. (2016) 30-band photometric redshifts. The color code is the same here as in Fig. 3, meaning that the shades of blue represent the *Euclid* selection, and the shades of gray represent the rest of the L15 sample. As in Fig. 3, undefined or negative point estimate values have been set to 0 in the plots.

Appendix D: CRPS plots

In Sect. 4.2, we present the mean and the median continuous ranked probability score (CRPS) for all the methods, specifically in Table 3. In Fig. D.1 we show the full distributions of CRPSs, for both the spectroscopic and the L15 samples. The spectroscopic sample CRPSs have their distribution weighted by the color-space weights (see Sect. 4.1).

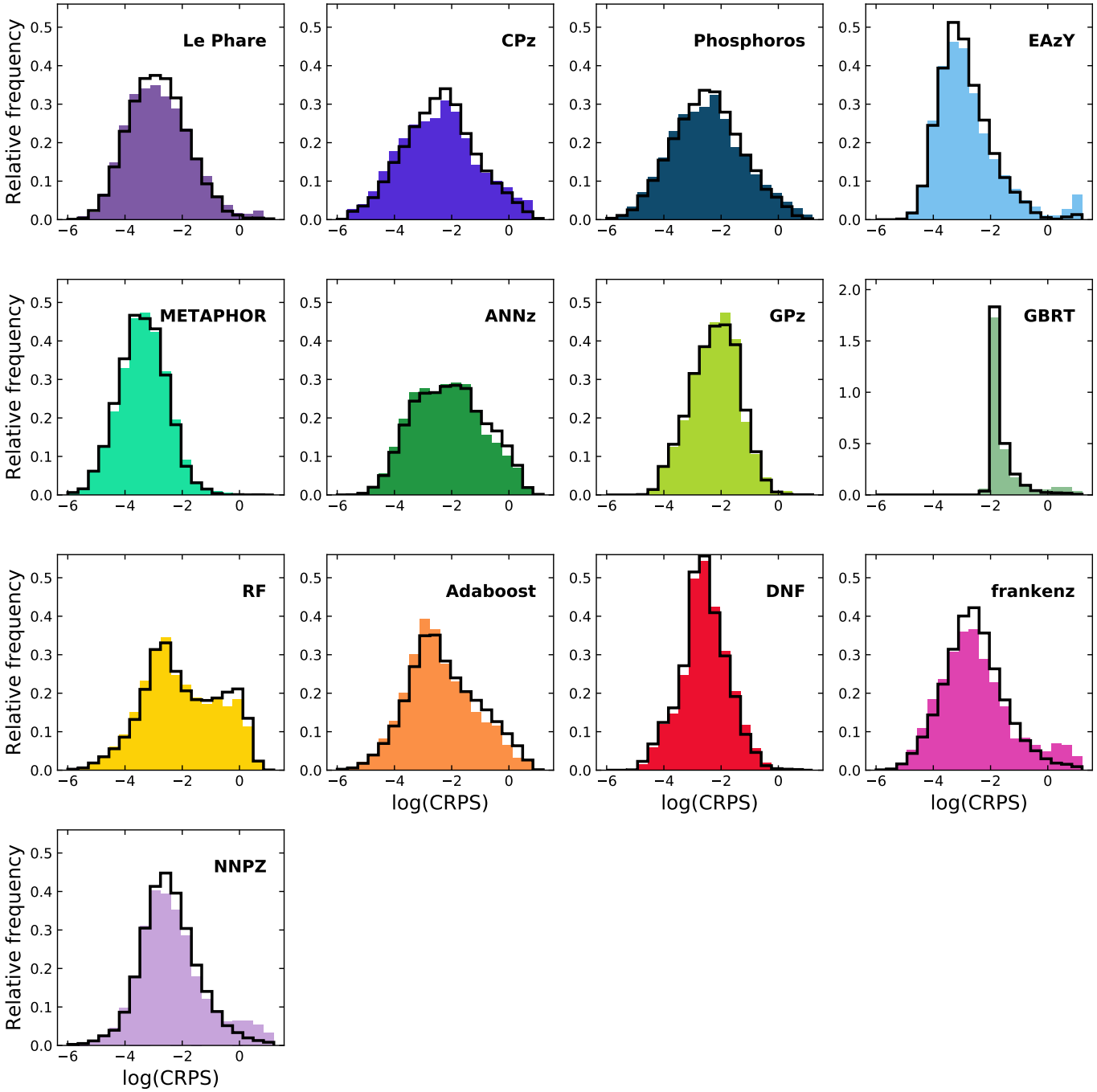

Fig. D.1. CRPS plots for all the methods. Colored histograms are the histograms of log(CRPS) for the weighted spectroscopic sample, while solid black lines are the histograms for the L15 sample.

Appendix E: Fraction of good sources per bin

In Sect. 5.3 we discuss which metrics we should consider to maximize the figure of merit of the weak-lensing signal. In Eq. (7), we correct the F_{0XX} metrics in all the bins by the square root of the fraction of sources appropriately attributed to the considered bin. Figure E.1 show this fraction in all the bins, for all the methods, using both the weighted spectroscopic sample (top panel) and the L15 sample (bottom panel). These values were used to compute the metrics presented in Fig. 9.

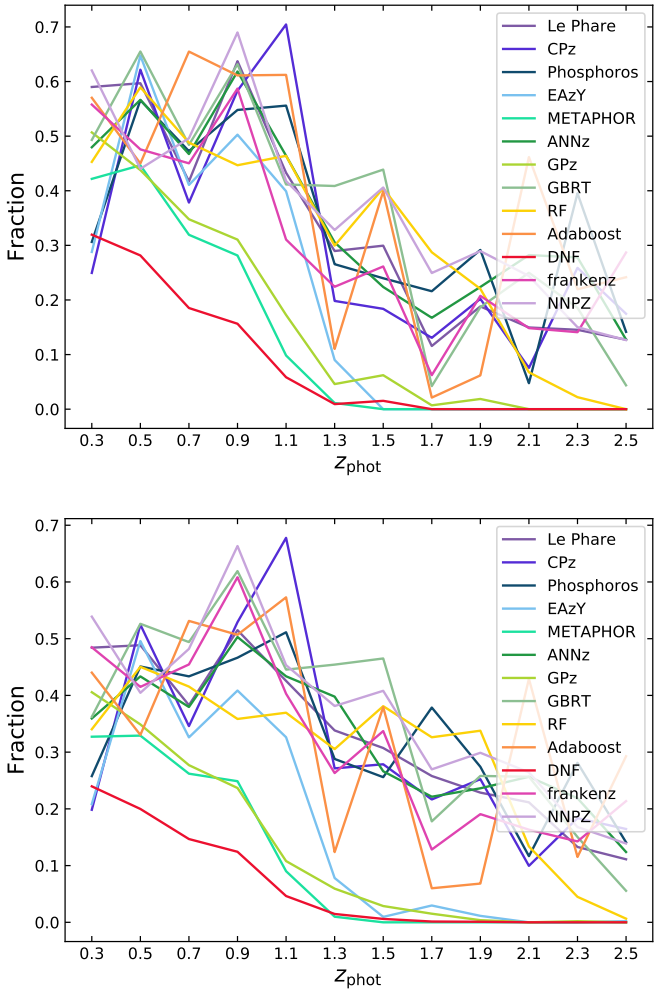

Fig. E.1. Fraction of sources per redshift bin that have both photo-z and true z belonging to the bin, compared to the number of sources with their spec-zs in the bin. *Top*: true z from GD the weighted spectroscopic sample. *Bottom*: true z from L15.