MNRAS **000**, 000–000 (0000) Preprint 18 December 2018 Compiled using MNRAS LaTeX style file v3.0

Implicit assumptions and their impact on photometric redshift PDF performance for LSST

S.J. Schmidt[1], A.I. Malz[2,3], J.Y.H. Soo[4], M. Brescia[5], S. Cavuoti[5,6], G. Longo[6], I.A. Almosallam[7,8], M.L. Graham[9], A.J. Connolly[9], E. Nourbakhsh[1], J. Cohen-Tanugi[10], H. Tranin[10], P.E. Freeman[11], K. Iyer[12], J.B. Kalmbach[13], E. Kovacs[14], A.B. Lee[11], C. Morrison[9], J. Newman[15], E. Nuss[10], T. Pospisil[11], M.J. Jarvis[16,17], R. Izbicki[18,19]

(LSST Dark Energy Science Collaboration)

[1] *Deptartment of Physics, University of California, One Shields Ave., Davis, CA, 95616, USA*
[2] *Center for Cosmology and Particle Physics, New York University, 726 Broadway, New York, 10003, USA*
[3] *Department of Physics, New York University, 726 Broadway, New York, 10003, USA*
[4] *Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[5] *INAF-Capodimonte Observatory, Salita Moiariello 16, I-80131, Napoli, Italy*
[6] *Department of Physics E. Pancini, University Federico II, via Cinthia 6, I-80126, Napoli, Italy*
[7] *King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia*
[8] *Information Engineering, Parks Road, Oxford, OX1 3PJ, UK*
[9] *Department of Astronomy, University of Washington, Box 351580, U.W., Seattle WA 98195, USA*
[10] *Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS, Montpellier, France*
[11] *Department of Statistics & Data Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*
[12] *Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854-8019 USA*
[13] *Department of Physics, University of Washington, Box 351560, Seattle, WA 98195, USA*
[14] *Argonne National Laboratory, Lemont, IL 60439, USA*
[15] *Department of Physics and Astronomy and PITT PACC, University of Pittsburgh, 3941 O'Hara St., Pittsburgh, PA 15260, USA*
[16] *Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK*
[17] *Department of Physics and Astronomy, University of the Western Cape, Bellville 7535, South Africa*
[18] *Department of Statistics, Federal University of Sao Carlos, Sao Carlos, Brazil*
[19] *External colaborator*

18 December 2018

ABSTRACT

In order to maximize scientific returns of current and upcoming galaxy surveys, the photometric redshift (photo-z) posterior distributions produced by redshift estimation codes must be accurate probability distribution functions (PDFs). However, the posteriors resulting from a number of current techniques are not, in general, consistent with each other, affected by implicit assumptions made by each code, and an optimal method for obtaining an accurate PDF estimate remains unclear. We present the results of an initial study by the Large Synoptic Survey Telescope Dark Energy Science Collaboration (LSST-DESC) testing twelve photo-z algorithms using complete and representative training data and evaluate multiple metrics to test how accurately the posteriors represent probability distributions. We observe several trends, including systematic biases and an overall over/under-prediction in the broadness of the PDFs in many of the codes which may be symptomatic of implementation problems or problems in underlying algorithm design. A careful accounting of all photo-z systematics will be necessary for the codes employed in upcoming analyses in order to achieve unbiased cosmological measurements.

Key words: galaxies: distances and redshifts – galaxies: statistics – methods: statistical

1 INTRODUCTION

Large-scale photometric galaxy surveys are entering a new era with currently or soon-to-be running Stage III and Stage IV dark energy experiments like the Dark Energy Survey (DES, Abbott et al. 2005), the Kilo-Degree Survey (KiDS, de Jong et al. 2013), Hyper Suprime-Cam (HSC) Survey (Aihara et al. 2018a,b), Large Synoptic Survey Telescope (LSST, Abell et al. 2009), Euclid (Laureijs et al. 2011), and Wide-Field Infrared Survey Telescope (WFIRST, Green et al. 2012). The move to imaging based surveys, rather than spectroscopic based, for cosmological measurements makes proper understanding of photometric redshifts ("photo-z's") of paramount importance, as cosmological distance measures for statistical samples are directly dependent on photo-z measurements.

The unprecedented sample size of LSST, expected to number several billion galaxies for the main cosmological sample, necessitates stringent constraints on photo-z accuracy if systematic errors are not to dominate the statistical errors. The LSST Science Requirements Document (SRD)[1] lists the individual galaxy photometric redshift goals for a magnitude limited sample with $i < 25$ as: root-mean-square error with a goal of $\sigma_z < 0.02(1+z)$; 3σ "catastrophic outlier" rate below 10%; bias below 0.003[2]. The LSST Dark Energy Science Collaboration (LSST-DESC) developed a separate Science Requirements Document (The LSST Dark Energy Science Collaboration et al. 2018), which forecasts the constraining power of five cosmological probes using somewhat conservative assumptions to define requirements on systematic errors for several measurements. These include even more stringent requirements on photometric redshift performance than those included in the LSST SRD, though most of the current LSST-DESC requirements are defined in terms of tomographic bin populations rather than on individual object redshifts. The tremendous size of LSST's galaxy catalogue will be enabled by its exceptional depth, pushing to fainter magnitudes and deeper imaging and including galaxies of lower luminosity and higher redshift than ever before. The inclusion of these populations introduce major physical degeneracies, for example the Lyman break/Balmer break degeneracy, that were not present in the populations covered in shallower large area surveys like the Sloan Digital Sky Survey (SDSS, York et al. 2000) and the Two Micron All Sky Survey (2MASS, Skrutskie et al. 2006). These issues are not unique to LSST, and are present in Stage III Dark Energy surveys; however, in order to meet the demanding error budgets of Stage IV projects such as LSST and LSST-DESC it will be necessary to fully characterize those degeneracies wherein multiple redshift solutions have comparable likelihood to per cent level accuracy.

There is often a desire to have a single valued "point-estimate" redshift for an individual galaxy. However, the complex, non-linear (and often non-unique) nature of the mapping between broad band fluxes and redshift means that a single value is unable to capture the full redshift information encoded in a galaxy's magnitudes. For example, a common point-estimate for a template-based method is taking the highest likelihood solution as the point photo-z. A single valued redshift ignores degenerate redshift solutions of lower probability, potentially biasing photometric redshift estimates both for individual galaxies and ensemble distributions. Storing more information is necessary, most often photo-z codes output the redshift probability density function, also often referred to as $p(z)$, describing the relative likelihood as a function of redshift. Early template methods such as Fernández-Soto et al. (1999) converted relative χ^2 values of template spectra to likelihoods to estimate $p(z)$. Soon after, codes such as Benítez (2000) added a Bayesian prior and output a posterior probability distribution. While many early machine learning based algorithms focused on a point-estimate, Firth et al. (2003) used a neural net with 1000 realizations scattered within the photometric errors to estimate a $p(z)$. As more groups began to employ photometric redshifts in their cosmological analyses, there was a realization that point-estimate photo-z's were inadequate for precision cosmology measurements (Mandelbaum et al. 2008). From around this point onward, most photo-z algorithms have attempted to implement some estimate of the overall redshift probability in their outputs, and some surveys began supplying a full $p(z)$ rather than a simple redshift point-estimate and error (e. g. de Jong et al. 2017).

For cosmological measurements, certain science cases require redshift information on individual objects, e. g. identification of host galaxy redshift for supernova classification, or identifying potential cluster membership. Other science cases seem to need only ensemble redshift information; for instance many current cosmic shear techniques require only the overall redshift distribution $N(z)$ for tomographic redshift samples. However, even such cases require individual object redshift estimates for portions of the analysis, for example in determining galaxy intrinsic alignments in weak lensing samples. In addition, recent data-driven techniques employing hierarchical Bayesian or Gaussian Process methods have emerged that calibrate redshift distributions using individual $p(z)$ estimates (e. g. Sánchez & Bernstein 2018). These methods assume that the $p(z)$ for each galaxy is an accurate PDF, and such methods break down if this assumption is invalid. Thus, even methods that seem to need only ensemble $N(z)$ may actually require accurate $p(z)$ in order to meet stringent survey requirements. Large photometric surveys such as LSST must develop algorithms that simultaneously meet the needs of all science cases. In order to meet these ambitious goals for photo-z accuracy, every aspect of photo-z estimation will have to be optimized: the algorithms employed, both template and machine-learning based (both in design and implementation); the spectroscopic data used as a training set for machine learning algorithms or to estimate template sets and train Bayesian priors; and probabilistic catalogue compression schemes that balance information retention against limited storage resources.

There are numerous techniques for deriving photo-z PDFs from photometry, yet no one method has yet been established as clearly superior. Quantitative comparisons of photo-z methods have been made before. The Photo-z Accuracy And Testing (PHAT, Hildebrandt et al. 2010) effort focused on point estimates derived from many photomet-

[1] available at `https://docushare.lsstcorp.org/docushare/dsweb/Get/LPM-17`

[2] Note that at the time the SRD was written, these goals were stated in terms of a photo-z point estimate for each galaxy, as was standard in many previous studies, while in this paper we emphasize the importance of using a full photo-z PDF.

ric bands. Rau et al. (2015) introduced a new method for improving redshift PDFs using an ordinal classification algorithm. DES compared several codes for point estimates and a subset with $p(z)$ information (Sánchez et al. 2014). A follow up up paper examined summary statistics of photo-z interim posteriors for tomographically binned galaxy subsamples (Bonnett et al. 2016).

This paper is distinguished by its focus on metrics of photo-z interim posteriors themselves and consideration of both classic and state-of-the-art photo-z algorithms, comparing the performance of several of the most widely employed codes, as well as some that have been developed only recently, on the basis of metrics appropriate for a probabilistic data product. The results presented here are a major focus of the Photometric Redshift working group of the LSST-DESC. This work is laid out in the Science Roadmap (SRM)[3] as one of the critical activities to be completed in preparation for dark energy science analysis on the first year LSST data. In this initial paper we focus on evaluating the performance of photometric redshift codes and PDF-based performance metrics in the presence of complete and representative training sets. Specific implementation choices in each code will influence the resultant posterior distributions, for example choice of prior parameterization in template-based codes, the bandwidth size chosen for machine learning based codes, or even the output format chosen for storing the PDF. We have attempted to minimize the impact of many of these factors when comparing codes, for example by using the same template set for all template-based codes, and using a training set that is drawn from the same underlying population as the test sample, to create a controlled environment in which to compare the photo-z PDFs derived from each method. We explore a number of performance metrics in this paper that test whether the posterior estimates are actual PDFs. Comparing the relative performance of the codes enables us to evaluate whether each code is using information in an optimal way, and may reveal enhancements in some codes and deficiencies in others, either in the fundamental algorithm, or in specific implementation. Identifying and fixing failure modes within codes may aid us in reaching the stringent photo-z performance goals set out for LSST. We note that these initial tests are a necessary requirement for photo-z codes that will be used in cosmological analyses; however, meeting these requirements is only the first stage in the process, and can be thought of as an initial test under near perfect conditions to test for problems before further complexities are added in future analyses.

The outline of the paper is as follows: in § 2 we present the simulated data set; in § 3 we describe the current generation codes employed in the paper; in § 4 we discuss the interpretation of photo-z PDFs in terms of metrics of accuracy; in § 5 we show our results and compare the performance of the codes; in § 6 we offer our conclusions and discuss future extensions of this work.

2 THE SIMULATION AND MOCK GALAXY CATALOG

In order to test the current generation codes, we employ an existing simulated galaxy catalogue. The simulation is completely catalogue-based, with no image construction or mock measurements made. We describe these in detail below.

2.1 Buzzard-v1.0 simulation

The BUZZARD-HIGHRES-V1.0 (De Rose et al., in prep; Wechsler et al., in prep) catalogue construction started with a dark matter only simulation. This N-body simulation contained 2048^3 particles in a 400 Mpc h^{-1} box. A set of time snapshots (with smoothing and interpolation between snapshots) were saved in order to construct a lightcone. Dark matter halos were identified using the ROCKSTAR software package (Behroozi et al. 2013). These dark matter halos were populated with galaxies with a stellar mass and absolute r-band magnitude in the SDSS system determined using a sub-halo abundance matching model constrained to match both projected two-point galaxy clustering statistics and an observed conditional stellar mass function (Reddick et al. 2013).

To assign an SED to each galaxy, the *Adding Density Dependent Spectral Energy Distributions* (ADDSEDS, deRose in prep.)[4] procedure was used. This consisted of training an empirical relation between absolute r-band magnitude, local galaxy density, and SED using a sample of $\sim 5\times10^5$ galaxies from the magnitude-limited Sloan Digital Sky Survey Data Release 6 Value Added Galaxy Catalog (Blanton et al. 2005). Each SDSS spectrum is fit with a sum of five SED components using the K-CORRECT v4_3? software package[5] (Blanton & Roweis 2007), thus each galaxy SED is parameterized as five weights for the basis SEDs. The distance to the spatial projected fifth-nearest neighbour was used as a proxy for local density in the SDSS training sample. For each simulated galaxy, a galaxy with similar absolute r-band magnitude and local galaxy density was chosen from the training set, and that training galaxy's SED was assigned to the simulated galaxy. This process is done in such a way as to preserve the colour-density relation of galaxy environment. Given the SED, absolute r-band magnitude and redshift, we computed apparent magnitudes in the six LSST filter passbands, *ugrizy*. We assigned magnitude errors in the six bands using the simple model described in Ivezić et al. (2008), assuming full 10-year depth observations had been completed. The number of total 30-second visits assumed when generating the photometric errors differs slightly from the fiducial numbers assumed for LSST: we assume 60 visits in u-band, 80 visits in g-band, 180 visits in r-band, 180 visits in i-band, 160 visits in z-band, and 160 visits in y-band. In the course of simulating Gaussian photometric errors, we add noise to objects fluxes, and some of these noisy fluxes will become negative in one or more bands. We call such negative fluxes "non-detections" and signify them with a placholder magnitude of 99.0 in the catalog. Thus, further mentions of "non-detections" refer to objects that would be "looked at but not seen" in multi-band

[3] Available at: `http://lsst-desc.org/sites/default/files/DESC_SRM_V1_1.pdf`

[4] `https://github.com/vipasu/addseds`

[5] `http://kcorrect.org`

forced photometry, and the photo-z codes will treat them as such. Only 2.0 per cent of our sample galaxies contain a photometric band with a non-detection, the vast majority of which are in the u-band.

2.1.1 *Selection of training and test sets*

The total initial catalogue covered 400 square degrees and contained 238 million galaxies to an apparent magnitude limit of $r = 29$ and spanning the redshift range $0 < z \leq 8.7$. In order for statistical errors not to dominate, we need less than one million galaxies in our sample. Several studies claim that only a few tens of thousands of spectra are necessary to calibrate photo-z surveys to Stage IV requirements (e. g. Bernstein & Huterer (2010), Masters et al. (2017)). Therefore, we aim for a final number of training galaxies between 3×10^4 and 5×10^4 in our sample. In order to reduce our sample to a reasonable size, we limit our dataset to a subset of ~ 16.8 square degrees selected from five separate spatial regions of the simulation. Systematic problems with galaxy colors above $z > 2$ were observed, so the catalogue was limited to include only galaxies in the redshift range $0 < z \leq 2.0$. A random subset of the the remaining galaxies was chosen, and placed at random into either a "training" set (10 per cent of the sample), for which the galaxies true redshifts will be supplied, or a "test" set (the remaining 90 per cent of the sample), for which each code will need to predict a redshift PDF for each galaxy. Finally, we restrict our analysis to a sample with an apparent i-band magnitude limit $i < 25.3$, which give a signal-to-noise $\sim$ 30 for most galaxies, a cut often referred to as the expected "LSST Gold Sample". This magnitude cut results in a training set with 44 404 galaxies and a test set containing 399 356 galaxies. All subsequent results will evaluate this "gold sample" test set. In order to blind results, initially redshifts were not revealed for the "test" set, and were only supplied for the training sample galaxies. This prevented code runners from tweaking results and fitting to the specific test set.

2.1.2 *Templates*

As mentioned in Section 2.1, the SEDs in the Buzzard simulation are drawn from an empirical set of SEDs taken from the SDSS DR6 NYU-VAGC, a sample of roughly $\sim 5 \times 10^5$ galaxies with spectra in SDSS. To determine a finite set of templates to use with template fitting codes we take the five SED weight coefficients for each of the galaxies in the SDSS sample and run a simple K-means clustering algorithm on this five dimensional space. Each dimension was normalized such that it spanned an interval $[0, 1]$. The K-means clusters partition the five-dimensional space of coefficients into Voronoi cells, spanning the space of coefficients in a way that properly reflects the underlying density in the coefficients. Thus, the resultant SEDs constructed using the cell centers as weight coefficients will provide a reasonable spanning SED set. An ad-hoc number of $K = 100$ was chosen and the 100 K-means centre positions are taken as the weights for the K-CORRECT SED components to construct one hundred template SEDs. These 100 templates were provided, and the templates were used by both BPZ and LEPHARE; however, because EAZY was designed and written to use the same five basis templates employed by K-CORRECT when constructing our mock galaxies, EAZY was run using linear combinations of these five templates rather than using the 100 discrete templates. The ability to fit for linear combinations of templates highlights an important implementation difference between similar photo-z codes.

2.1.3 *Limitations*

For our initial investigation of photometric redshift codes, we begin with a data set that is somewhat idealized, and does not contain all of the complicating factors present in real data. In several cases, the simplification is done with a purpose, with potentially confounding effects excluded in order to better isolate the differences between current-generation photo-z codes and their causes. We list several of the simulation limitations in this section. As the simulation is catalogue-based, no image level effects, such as photometric measurement effects, object blending, contamination from sky background (Zodiacal light, scattered light, etc...), lensing magnification, or Galactic reddening are included. No stars are included in the catalogue, nor are the effects of AGN. As all SEDs are constructed from only five basis templates, properties of the galaxy population will be restricted to follow linear combinations of the characteristics of the five basis templates, so certain non-linear features, for example the full range of emission line fluxes relative to the continuum, will not be included in the model galaxy population. Moreover, the linear combinations of templates are modeled on the $\sim 5 \times 10^5$ SDSS galaxies discussed in Section 2.1, and thus only galaxies that resemble those spectroscopically observed by the SDSS will be included in the sample. No additional dust reddening intrinsic to the host galaxy is included, the only approximation of dust extinction comes in the form of dust encoded in the five basis SEDs via the training set used to create the basis templates. Simple linear combinations of these basis templates will, once again, not explore the full range of realistic dust extinction observed in galaxy populations. While these idealized conditions limit the realism of our galaxy population, some are also by design. We aim to test the photo-z codes at a very basic level, and a simplified model assures that differences in results seen between the codes are due to fundamental differences in their underlying assumptions and implementation details, rather than more nuanced properties.

3 METHODS

Here we outline the photo-z PDF codes tested in this study. In total, twelve distinct codes are tested. This sample is not comprehensive, codes were chosen based on the expertise available within the group; however, those chosen do cover a broad range of the current-generation methods used in the field. Both template-based and machine learning approaches are included and each are described separately in Secs. 3.1 and 3.2 respectively. The list of codes are summarized in Table. 1. All code runners were asked to output redshift posterior estimates on 200 linear-spaced bins between redshifts 0 and 2.

The questions that must be answered for each code are:

Table 1. List of photo-z codes featured in this study. ML here means machine learning.

Code	Type	Paper	Website
BPz	template	Benítez (2000)	http://www.stsci.edu/~dcoe/BPZ/
EAZY	template	Brammer et al. (2008)	https://github.com/gbrammer/eazy-photoz
LePhare	template	Arnouts et al. (1999)	http://www.cfht.hawaii.edu/~arnouts/lephare.html
ANNz2	ML	Sadeh et al. (2016)	https://github.com/IftachSadeh/ANNZ
Delight	ML/template	Leistedt & Hogg (2017)	https://github.com/ixkael/Delight
FlexZBoost	ML	Izbicki & Lee (2017)	https://github.com/tpospisi/flexcode; https://github.com/rizbicki/FlexCoDE
GPz	ML	Almosallam et al. (2016b)	https://github.com/OxfordML/GPz
METAPhoR	ML	Cavuoti et al. (2017)	http://dame.dsf.unina.it
CMNN	ML	Graham et al. (2018)	-
SkyNet	ML	Graff et al. (2014)	http://ccpforge.cse.rl.ac.uk/gf/project/skynet/
TPZ	ML	Carrasco Kind & Brunner (2013)	https://github.com/mgckind/MLZ
trainZ	N/A	See Section 3.3	

what unique features are included in the specific implementation that influence the output $p(z)$. What form of validation was performed with the training data, how were photometric uncertainties employed in the analysis, how were negative fluxes treated, what specific prior form was employed (for template based codes), or what specific machine learning architecture was used (for ML codes)?

3.1 Template-based Approaches

We test three publicly available and commonly used template-based codes: BPZ, EAZY, and LePhare. All three codes follow the standard procedure for template-based redshift estimation: calculate model fluxes for a set of template spectral energy distributions (SEDs) on a grid of redshift values and calculating a χ^2 merit function using the observed and model fluxes:

$$\chi^2(z,T,A) = \sum_i^{N_{\rm filt}} \left(\frac{F^i_{\rm obs} A \times F^i_{\rm pred}(T,z)}{\sigma^i_{\rm obs}} \right)^2 \quad (1)$$

where A is a normalization factor, $F^i_{\rm pred}(T,z)$ is the flux predicted for a template T at redshift z. $F^i_{\rm obs}$ is the observed flux in a given band i and $\sigma^i_{\rm obs}$ is the observed flux error. $N_{\rm tot}$ is the total number of filters, in our case the six $ugrizy$ LSST filters. Specific implementation details of each code, e. g. prior form and implementation, are described below.

3.1.1 BPZ

BPZ[6] (Bayesian Photometric Redshift, Benítez 2000) is a template-based photo-z code that compares the expected colors (C) calculated for a set of spectral energy distribution (SED) types/templates (T) to the observed colors to calculate the likelihood of observing colors at each redshift for each type, $p(C|z,T)$. The likelihoods at each redshift are related to the χ^2 in Equation 1 by the simple form: likelihood $\propto e^{-\chi^2/2}$. The code employs an empirically determined Bayesian prior in apparent magnitude (m_0) and SED-type. Assuming that the SED-types are spanning and exclusive, we can determine the redshift posterior $p(z|C,m_0)$ by marginalizing over all SED-types with a simple sum (Eq. 3 from Benítez 2000):

$$p(z|C,m_0) \propto \sum_T p(z,T|m_0)\, p(C|z,T) \quad (2)$$

where the first term on the right-hand side is the Bayesian prior and the second term is the traditional likelihood. The prior is assumed to have the form: $p(z,T|m_0) = p(T|m_0)\, p(z|T,m_0)$, i.e. it parameterizes the prior as an evolving type fraction with apparent magnitude, combined with a prior on the expected redshift probability distribution as a function of both apparent magnitude and SED-type.

In this paper we use BPZ v 1.99.3. The template set employed here is the set of 100 discrete SEDs described in Section 2.1.2 To keep the number of free parameters to a manageable level the SEDs in the training set are sorted by the rest-frame u-g colour and split into three "broad" SED classes, equivalent to the `E`, `Sp` and `Im/SB` types in Benítez (2000). We assume the same functional form for the Bayesian priors as used by Benítez (2000), and utilize the training-set galaxies with known SED-type, redshift, and apparent magnitude to determine the type fractions and the best fit for the eleven free parameters of the prior. For galaxies that are not detected in a measured band, the placeholder value is replaced with an estimate of the one σ detection limit in that particular band, i. e. a value close to the estimated sky noise threshold. The type-marginalized $p(z)$ is generated by setting the parameter `PROBS_LITE=TRUE` in the BPZ parameter file.

3.1.2 EAZY

EAZY[7] (Easy and Accurate Photometric Redshifts from Yale, Brammer et al. 2008) is a template-based photo-z code that includes several features that extend the basic χ^2 fit used in many template codes. The code can fit the observed photometry with SEDs created from a linear combination of a set of templates at each redshift, and the best-fit SED is found by simultaneously fitting one, two or all of the templates by minimizing χ^2. The minimized $\chi^2(z)$ is

[6] http://www.stsci.edu/~dcoe/BPZ/

[7] https://github.com/gbrammer/eazy-photoz

then combined with an apparent magnitude prior to obtain the posterior redshift probability distribution. On examination of the source code for EAZY, it appears that rather than marginalizing across all templates in the χ^2 calculation, EAZY takes only the minimum value of χ^2 at each redshift. This improper marginalization does not lead to the correct posterior distribution, an implementation issue that will need to be addressed in the future. EAZY can also account for the uncertainties in the templates by adding an empirically derived template error in quadrature as a function of redshift to the flux errors.

In this paper we use the all-templates mode, which fits the photometric data with a linear combination of the five basis templates. We employed the 5 basis templates described in Section 2.1, and set the template error to zero since these same templates were used to produce the simulated catalog photometry. The likelihoods include the application of a type-independent apparent magnitude prior estimated from the training data.

3.1.3 LePhare

LePhare[8](Photometric Analysis for Redshift Estimate, Arnouts et al. 1999; Ilbert et al. 2006) is a photo-z reconstruction code based on a χ^2 template-fitting procedure. The observed colors are matched with the colours predicted from a set of spectral energy distribution (SED) which can be either synthetic or based on a semi-empirical approach.

Each SED is convolved with the simulated LSST filter transmission curves (accounting for instrument efficiency). The computed photo-z is then the value that minimizes the merit function $\chi^2(z, T, A)$ from Arnouts et al. (1999), and given in Equation 1.

In this paper we use LePhare v 2.2. The set of templates used for fitting the photo-z's are the 100 discrete Buzzard SED templates as described in section 2.1.2, and the full $p(z)$ corresponds to the likelihoods calculated at each point on our z-grid.

3.2 Training-based Codes

The training-based codes use a variety of algorithms in order to estimate $p(z)$, specifics of each implementation are described in the subsections. Some aspects of data treatment were left to the individual code runners, for example, whether/how to split the available data with known redshifts into separate training and validation sets. Another key difference is the treatment of non-detections in one or more bands. Some codes choose to ignore a band, others replace the value with either an estimate for the detection limit, the mean of other values in the training set, or another default value. There are varying conventions among training-based codes for treatment of non-detections, and no one prescription dominates in the photo-z literature. The specific choices for each code affect the results, and contribute to the implicit prior influencing their output. However, we remind the reader that only 2.0 per cent of our sample has non-detections, almost exclusively in the u-band, and thus should not dominate the code performance differences.

[8] http://www.cfht.hawaii.edu/~arnouts/lephare.html

3.2.1 ANNz2

ANNz2[9] (Sadeh et al. 2016) is a software package that has the ability to employ several machine learning algorithms, including artificial neural networks (ANN), boosted decision tree (BDT) and k-nearest neighbour (KNN). Using the Toolkit for Multivariate Data Analysis (TMVA) with ROOT[10], it can either run a single machine learning algorithm and return results, or it can run multiple algorithms simultaneously and output photo-z's as a weighted combination of the different algorithms. In this study, only ANNs were employed. The redshift PDFs are produced by running an ensemble set of ANNs, each with different random seeds used in initialization of input parameters for training. Uncertainties for each method are estimated from a KNN-uncertainty estimator (Oyaizu et al. 2008). The final PDF can either be the “best” of the candidate PDFs, or a weighted average of the PDFs based on their error estimates for each of the ensemble members.

In this study, ANNz2 v. 2.0.4 was used. A set of 5 ANNs with architecture 6 : 12 : 12 : 1 (6 *ugrizy* inputs, 2 hidden layers with 12 nodes each, and 1 output) with different random seeds are used during each training. Half of the training set is used as a validation set to prevent overtraining. All training objects are set to have detected magnitudes, however the non-detections (mag= 99) in the testing set are replaced with the mean of that particular band.

3.2.2 Colour-Matched Nearest-Neighbours

The nearest-neighbours colour-matching photometric redshift estimator (CMNN) is presented in Graham et al. (2018, hereafter G18). This method uses a training set of galaxies with known redshifts that has equivalent or better photometry as the test set in terms of quality and filter coverage. For each galaxy in the test set we identify a colour-matched subset of training galaxies. This subset is identified by first calculating the Mahalanobis distance D_M in colour-space between the test galaxy and all training-set galaxies:

$$D_{\rm M} = \sum^{N_{\rm colours}} \frac{(c_{\rm train} - c_{\rm test})^2}{(\delta c_{\rm test})^2} \quad (3)$$

where c colour, $\delta c_{\rm test}$ is the measurement error on the colour, and $N_{\rm colours}$ is the total number of colors (i.e., in our case u-g, g-r, r-i, i-z, and z-y). Then, we choose a threshold value for D_M that defines the colour match set based on a set value of the percent point function (PPF): for example, for $N_{\rm dof} = 5$, choosing a vPPF$= 0.95$, 95 per cent of all training galaxies consistent with the test galaxy will have $D_M < 11.07$ (where $N_{\rm dof}$ is the number of degrees of freedom, in this case the number of colours). If a galaxy had a non-detection in a band, that band was dropped and $N_{\rm dof}$ was reduced by one in the colour-matching space. For a given test galaxy, the $p(z)$ is the normalized distribution of the true catalogue redshifts of this colour-matched subset of training galaxies.

We have applied the nearest-neighbours colour-matching photometric redshift estimator described in G18

[9] https://github.com/IftachSadeh/ANNZ

[10] http://tmva.sourceforge.net/

to the simulated data. Compared to its application in G18, there are some minor differences in the application of this estimator to the Buzzard catalogue. First, we do not impose non-detections on galaxies with a magnitude fainter than the expected LSST 10-year limiting magnitude or bright enough to saturate with LSST: *all* of the photometry for all the galaxies in the test and training sets are used for this experiment. Second, as in G18 we do apply an initial cut in colour to the training set before calculating the Mahalanobis distance in order to accelerate processing, and also use a magnitude pseudo-prior to improve photo-z estimates, but for both we have used different cut-off values that are appropriate for the Buzzard galaxies' colours and magnitudes. Third, we set different parameters for the identification of the colour-matched subset of training galaxies and the selection of a photometric redshift estimate. In G18 we used a PPF value of 0.68 to identify the colour-matched subset of training galaxies and used the redshift of nearest neighbour in colour-space as the photo-z estimate. These choices work well when the desire is to obtain accurate photo-z estimates for most test-set galaxies, but does not return an accurate $p(z)$ in all cases – especially for galaxies that are bright and/or have few matches in colour-space. Since an accurate estimate of $p(z)$ is desired for this work we make several changes to our implementation of the CMNN photo-z estimator. We continue to use a PPF = 0.95 to generate the subset of colour-matched training galaxies, but weight them by the inverse of their Mahalanobis distance. This weighting maintains some of the accuracy that was previously achieved by simply using the nearest neighbour in colour-space. We then use the weights to create the $p(z)$ instead of having the redshift of each colour-matched training-set galaxy count equally. To obtain a robust estimate of the $p(z)$ for galaxies with a small number of colour-matched training set galaxies, when this number is less than 20 the nearest 20 neighbours in colour-space are used instead, and we convolve the $p(z)$ with a Gaussian with a standard deviation of:

$$\sigma = \sigma_{\rm train}\sqrt{(\mathrm{PPF}_{20}/0.95)^2 - 1} \tag{4}$$

to appropriately broaden it so that the $p(z)$ for these test galaxies represents the enlarged PPF value associated with it. Overall, these three changes will yield less precise photo-z estimates compared to those presented in G18, but they will all have significantly more accurate estimates of the $p(z)$, particularly for the brightest test galaxies. This is sufficient for this work because, as described in G18, the goal of the CMNN photo-z estimator was never to provide the "best" (or even competitive) estimates in the first place, given its reliance on a deep training set, but rather to provide a means for direct comparisons between LSST photometric quality and photo-z estimates. With this work we show how the input parameters should be set in order to return accurate $p(z)$ estimates in addition to point value estimates.

3.2.3 Delight

DELIGHT[11] (Leistedt & Hogg 2017) infers photo-z's by using a data-driven model of latent SEDs and a physical model of photometric fluxes as a function of redshift. Delight models the underlying latent SEDs as a linear combination of a set of pre-defined template SEDs, plus zero mean Gaussian processes with factorized kernels. Generally, machine learning methods rely on representative training data with similar band passes, while template based methods rely on a complete library of templates based on physical models constructed. DELIGHT is constructed in attempt to combine the advantages and eliminate the disadvantages of both template-based and machine learning algorithms: it constructs a large collection of latent SED templates (or physical flux-redshift models) from training data, with a template SED library as a guide to the learning of the model. The advantage of DELIGHT is that it neither needs representative training data in the same photometric bands, nor does it need detailed galaxy SED models to work.

This conceptually novel approach is done by using Gaussian processes operating in flux-redshift space. The posterior distribution on the redshift of a target galaxy is obtained via a pairwise comparison with training galaxies,

$$p(z|\hat{\mathbf{F}}) \approx \sum_i p(\hat{\mathbf{F}}|z,t_i)\, p(z|t_i) p(t_i), \tag{5}$$

where $p(z|t_i)p(t_i)$ captures prior information about the redshift distributions and abundances of the galaxies, with t_i denoting the galaxy template; while $p(\hat{\mathbf{F}}|z,t_i)$ is the posterior of noisy flux $\hat{\mathbf{F}}$ at redshift z. For each training-target pair, $p(\hat{\mathbf{F}}|z,t_i)$ is evaluated as follows:

$$p(\hat{\mathbf{F}}|z,t_i) = \int p(\hat{\mathbf{F}}|\mathbf{F})\, p(\mathbf{F}|z,z_i,\hat{\mathbf{F}}_i)\, d\mathbf{F}, \tag{6}$$

where $p(\hat{\mathbf{F}}|\mathbf{F})$ is the likelihood function, it compares the noisy real flux $\hat{\mathbf{F}}$ with the noiseless flux $\mathbf{F}$ obtained from the linear combination of template models, carefully constructed to account for model uncertainties and different normalization of the same SED; while $p(\mathbf{F}|z,z_i,\hat{\mathbf{F}}_i)$ is the prediction of flux at a different redshift z with respect to the training object with redshift z_i and flux $\hat{\mathbf{F}}_i$. Eq. 6 is essentially the probability that the training and the target galaxies having the same SED but at a different redshift. The flux prediction $p(\mathbf{F}|z,z_i,\hat{\mathbf{F}}_i)$ of the training galaxy at redshift z is modeled via a Gaussian process,

$$F_b \sim \mathcal{GP}\left(\mu^F, k^F\right), \tag{7}$$

with mean function μ^F and kernel k^F, both imposed to capture expected correlations resulting from the known underlying physics (i.e., fluxes resulting from observing SEDs through filter response, and the SEDs being redshifted). The reader should refer to Leistedt & Hogg (2017) for further details.

In this study, all 100 ordered Buzzard templates, as described in Section 2.1.2, were used in DELIGHT, and the Gaussian process was trained using the provided training sample. Photometric uncertainties from the inputs are propagated into the code, while non-detections for each band are set to the mean of the respective bands. The default settings of DELIGHT were used, with the exception that the PDF bins were set to be linearly-spaced rather than logarithmic. In this study a flat prior in magnitude/type is assumed.

[11] https://github.com/ixkael/Delight

3.2.4 FlexZBoost

FLEXZBOOST[12] (Izbicki & Lee 2017) is a particular realization of `FlexCode`, which is a general-purpose methodology for converting any conditional mean point estimator of z to a conditional *density* estimator $f(z|\mathbf{x})$, where $\mathbf{x}$ here represents our photometric covariates and errors.[13] The key idea is to expand the unknown function $f(z|\mathbf{x})$ in an orthonormal basis $\{\phi_i(z)\}_i$:

$$f(z|\mathbf{x}) = \sum_i \beta_i(\mathbf{x})\phi_i(z). \tag{8}$$

By the orthogonality property, the expansion coefficients are just conditional means

$$\beta_i(\mathbf{x}) = \mathbb{E}\left[\phi_i(z)|\mathbf{x}\right] \equiv \int f(z|\mathbf{x})\phi_i(z)dz. \tag{9}$$

These coefficients can easily be estimated from data by regression since $\mathbb{E}\left[\phi_i(z)|\mathbf{x}\right]$ is the regression of $\phi_i(z)$ (a transformation of Z) on X.

In this paper, we use XGBOOST (Chen & Guestrin 2016) for the regression part; it should however be noted that FLEXCODE-RF (also on GitHub), based on Random Forests, generally performs better for smaller data sets. As our basis, we choose a standard Fourier basis. There are two tuning parameters in our $p(z)$ estimate: (i) the number of terms, I, in the series expansion in Eq. 8, and (ii) an exponent α that we use to sharpen the computed density estimates $\widehat{f}(z|\mathbf{x})$, according to $\widetilde{f}(z|\mathbf{x}) \propto \widehat{f}(z|\mathbf{x})^\alpha$. We reserve 15% of the training set data as a validation set, and choose both I and α in an automated way by minimizing the weighted L_2-loss function (Eq. 5 in Izbicki & Lee 2017) on the validation set. While the native storage format for FLEXCODE encodes the PDF using the coefficients shown in Equation 9, to match the output format requested of other codes we discretize our final estimates into 200 bins linearly spaced in $0 < z < 2$.

3.2.5 GPz

GPZ[14] (Almosallam et al. 2016a,b) is a sparse Gaussian process based code, a scalable approximation of full Gaussian Processes (Rasmussen & Williams 2006), with the added feature of being able to produce input-dependent variance estimations (heteroscedastic noise). The model assumes that the probability of the output y, the redshift, given the input x, the photometry, is $p(z|x) = \mathcal{N}\left(z|\mu(x), \sigma(x)^2\right)$. The mean function, $\mu(x)$, and the variance function $\sigma(x)^2$ are both linear combinations of basis functions that take the following form:

$$f(x) = \sum_{i=1}^{m} \phi_i(x)w_i, \tag{10}$$

where $\{\phi_i(x)\}_{i=1}^m$ and $\{w_i\}_{i=1}^m$ are sets of m basis functions and their associated weights respectively. Basis function models (BFM), for specific classes of basis functions such as the sigmoid or the squared exponential, have the advantage of being universal approximators, i.e. there exist a function of that form that can approximate any function, with mild assumptions, to any desired degree of accuracy. The details on how to learn the parameters of the model and the hyper-parameters of the basis functions are described in Almosallam et al. (2016b).

A unique feature in GPZ, is that the variance estimate is composed of two terms each quantifying a different source of uncertainty. One term (the model uncertainty) reflects how much of the uncertainty is due to lack of training samples at the location of interest, whereas the second term (the noise uncertainty) reflects how much of the uncertainty is caused from observing many noisy samples at that location. Thus, the predictive variance can determine whether we need more representative samples or more precise samples for any particular location in the input space. GPZ can also emphasize the importance of some samples as weights. This weight can be for example $|z_{\rm spec} - z_{\rm phot}|/(1 + z_{\rm spec})$ to target the desired objective of minimizing the normalized redshift error or as a function of their probability in the test set relative to the training set in order to pressure the model to better fit samples that are rare in the training set but are expected to be abundant during testing.

The data is prepared for GPZ by taking the log of the magnitude errors, decorrelating the data set using PCA and imputing any missing magnitude values using a simple linear model that estimates the missing magnitudes given the observed ones. The log transformation helps to smooth the long tail distribution of the magnitude errors, which is more stable numerically and makes the optimization process unconstrained. The missing values are imputed by computing the mean of the training set μ and its covariance Σ, then we use the following equation to estimate the missing values from the observed ones

$$x_u = \mu_u + \Sigma_{uo}\Sigma_{oo}^{-1}(x_o - \mu_o), \tag{11}$$

where the subscript o in x_o indexes the *observed* part of the input x, whereas the subscript u indexes the *unobserved* set (similarly for μ and Σ). This is the optimal expected value of the unobserved variables given the observed ones if the distribution is jointly Gaussian, note that if the variables are independent, i.e. $\Sigma_{uo} = 0$, this will reduce to a simple average predictor. We use the Variable Covariance (VC) option in GPZ with 200 basis functions after we note that there is no significant increase in the performance on the validation set (using 80%-20% training-validation split) and with no cost-sensitive learning applied.

3.2.6 METAPhoR

METAPHOR (Machine-learning Estimation Tool for Accurate Photometric Redshifts, Cavuoti et al. 2017) is a pipeline designed to provide photo-z point estimates and a reliable PDF for machine learning (ML) based techniques. It includes pre- and post-processing phases, hosting a photo-z prediction engine based on the Multi Layer Perceptron with Quasi Newton Algorithm (MLPQNA).

METAPHOR includes data modules for pre-processing, photo-z estimation, and PDF estimation, and post-processing. The pre-processing includes a model for pertur-

[12] https://github.com/tpospisi/flexcode; https://github.com/rizbicki/FlexCoDE
[13] Instead of $p(z)$, we use the notation $f(z|\mathbf{x})$ to explicitly show the dependence on $\mathbf{x}$.
[14] https://github.com/OxfordML/GPz

bation of the photometry that is employed in calculating the PDF of the photo-z estimation errors. The photometric perturbation is defined as:

$$m_{ij} = m_{ij} + \alpha_i F_{ij} u_{\mu=0,\sigma=1} \tag{12}$$

where α_i is a user selected multiplicative constant (useful in case of multi-survey photometry), $u_{\mu=0,\sigma=1}$ is a random value from the standard normal distribution and F_{ij} is a bimodal function (a constant function + polynomial fitting of the mean magnitude errors on the binned bands), heuristically tuned in such a way that the constant component is the threshold under which the polynomial function is considered too low to provide a significant noise contribution to the photometry perturbation.

As main prerogative, METAPHoR is able to provide a PDF for ML methods by taking into account the photometric errors provided with data, by running N trainings on the same training set, or M trainings on M different random extractions from the KB. The different test sets, used to produce the PDF, are thus obtained by introducing a proper perturbation, parametrized from the photometric error distribution in each band, on the photometric data populating the original test set (Brescia et al. 2018). For the present work since it was required to produce a redshift (and a PDF) for each object of the test set we decided to apply a hierarchical kNN to replace the missing detections with values based on their neighbors. The reliability of PDFs and point estimation is lower. No cross validation has been used.

3.2.7 *SkyNet*

SkyNet[15] (Graff et al. 2014) is a publicly available neural network software, based on a 2nd order conjugate gradient optimization scheme (see Graff et al. 2014, for further details).

The neural network is configured as a standard multilayer perceptron with three hidden layers and one input layer with 12 nodes (the 6 magnitudes and their errors). The classifier is laid out such that the hidden layers have 20:40:40 nodes each, all rectified linear units, and the output layer has 200 nodes (corresponding to 200 bins for the PDF) activated with a "softmax" function so that they automatically sum to 1. While previous implementations of the code, such as Sánchez et al. (2014) and Bonnett (2015) (see Appendix C.3), implement a "sliding bin" smoothing, no such procedure was used in this study.

To avoid over-fitting, a 30 per cent fraction of the training set is used as validation, and the training is stopped as soon as the error rate begins to increase in the validation set. The weights are randomly initialized based on normal sampling. The error function is a standard chi-square function for the regressor, and a cross-entropy function for the classifier. Finally, the data are all whitened before processing, with magnitudes pegged to (45,45,40,35,42,42) and their errors pegged to (20,20,10,5,15,15) for *ugrizy* filters, respectively.

3.2.8 *TPZ*

TPZ[16] (Trees for Photo-z, Carrasco Kind & Brunner 2013; Carrasco Kind & Brunner 2014) is a parallel machine learning algorithm that generates photometric redshift PDFs using prediction trees and random forest techniques. The code recursively splits the input data (i. e. the training sample), into two branches, one after another, until a terminal leaf is created that meets a termination criterion (e. g. a minimum leaf size or a variance threshold). Bootstrap samples from the training data and associated errors are used to build a set of prediction trees. In order to minimize correlation between the trees, the data is divided in such a way that the highest information gain among the random subsample of features is obtained at every point. The regions in each terminal leaf node corresponds to a specific subsample of the entire data that possesses similar properties.

The training data is examined before running TPZ. Since TPZ does not handle non-detections (magnitudes flagged as 99.0), we replace these values with an approximation of the 1σ detection threshold, i. e. a signal to noise ratio of 1 in terms of magnitude uncertainty using the equation $dm = 2.5 \log(1 + N/S)$ where $dm \sim 0.7526$ *mag* for $N/S = 1$. That is, for each band, we replace the non-detection with the magnitude corresponding to the error of 0.7526 from the error model forecasted for 10-year LSST data. The Out-of-Bag (Breiman et al. 1984; Carrasco Kind & Brunner 2013) cross-validation technique is used within TPZ to evaluate its predictive validity and determine the relative importance of the different input attributes. We employed this information to calibrate our algorithm.

In the present work, the LSST magnitudes u, g, r, i and colours $u-g$, $g-r$, $r-i$, $i-z$, $z-y$ and their associated errors are used in the process of growing 100 trees with a minimum leaf size of 5 (the z and y magnitudes did not show significant correlation with the redshift in our cross-validation, so we did not use them when constructing our trees). We partitioned our redshift space into 200 bins and smoothed each individual PDF with a smoothing scale of twice the bin size.

3.3 Simple Ensemble Estimator

In addition to the main photo-z algorithms described above we also include a very simple method as a pathological example. For TRAINZ, as we will we call this simple estimator, we well define $p(z)$ as simply:

$$p(z) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} z_{train} \tag{13}$$

That is, we simply set the redshift PDF of every galaxy equal to the normalized $N(z)$ of the training sample. This estimator is essentially a k nearest-neighbour estimator with k equal to the number of galaxies in the training sample. As the training sample is drawn from the same underlying distribution as the test sample, modulo small deviations due to sample size, the quantiles of the training and test distributions should be identical, modulo fluctuations due to

[15] http://ccpforge.cse.rl.ac.uk/gf/project/skynet/

[16] https://github.com/mgckind/MLZ

finite sample size. This is a wildly unrealistic estimator, as it assigns all galaxies, no matter their apparent magnitude, colour, or true redshift, the same redshift PDF, and is thus uninformative at the level of individual object redshifts, but is designed to perform very well for the ensemble of all objects. If the training set was not representative, this estimator would produce biased results, and any attempts to break up the sample into tomographic bins will fail, as every galaxy has an identical $p(z)$. We will discuss this method and cautions relative to metrics in Section 5.3.

4 METRICS FOR QUANTIFYING PDF COMPARISONS

The overloaded "$p(z)$" is a widespread abuse of notation; we would like the outputs of photo-z PDF codes to be interpretable as probabilities. Obviously photo-z PDFs must not take negative values and must integrate to unity over the range of possible redshifts. Additionally, an estimator derived by method H for the photo-z PDF of galaxy i must be understood as a posterior probability distribution

$$\hat{p}_j(z_i) = p(z|d_i, I_D, I_H), \tag{14}$$

conditioned not only on the photometric data d_i for that galaxy but also on parameters encompassing a number of things that will differ depending on the method H used to produce it, namely the assumptions I_H necessary for the method to be valid and any inputs I_D it takes as prior information, such as a template library or training set. Because of this, direct comparison of photo-z PDFs produced by different methods is in some sense impossible; even if they share the same prior information I_D, by definition they cannot be conditioned on the same assumptions I_H, otherwise they would not be distinct methods at all.

In this study, we isolate the differences in prior information specific to each method by using a single training set I_D^{ML} for all machine learning-based codes and a single template library I_D^T for all template-based codes, and these sets of prior information are carefully constructed to be representative and complete, we have $I_D^{ML} \equiv I_D^T$ for every method H. Thus, we are saying

$$\frac{\hat{p}_{i,H}(z)}{\hat{p}_{i,H'}(z)} \approx \frac{p(z|d_i, I_H)}{p(z|d_i, I_{H'})}, \tag{15}$$

meaning that we assume comparisons of $\hat{p}_{i,H}(z)$ isolate the effect of the method used to obtain the estimator, which should make examination of differences caused by specifics of the method implementations easier to isolate.

As mentioned previously, there are cosmology probes that require knowledge of individual galaxy's photo-z PDF, for example galaxy intrinsic alignment studies, strong lensing foreground shear prediction, and photometric supernova classification, and others that require only knowledge of the ensemble redshift distribution, $N(z)$. Due to the paucity of principled techniques for using and validating individual galaxy photo-z PDFs, there have been few alternatives to the common practice of reducing photo-z PDFs to point estimates when evaluating and comparing photo-z code performance in the literature. Though this practice should not be encouraged, we also calculate traditional metrics based on the most common point estimators derived from photo-z PDFs. Those seeking to establish a connection to traditional ways of thinking about redshift estimation may consult the Appendix for these results.

There are a number of metrics that can be used to test the accuracy of a photo-z posterior as an estimator of a true photo-z posterior if it is known. Even for simulated data, the true photo-z PDF is in general not accessible, as rather than drawing photometric galaxy redshifts from a true photo-z likelihood distributions for a given set of photometry, mock cataloges are instead generated by populating N-body simulation based light cones with photometry generated from SED models, so the relation between redshift and photometric properties is indirect. Furthermore, only limited applications of photo-z PDFs that could be used as the basis for a metric have been presented in the literature.

The most popular application of photo-z PDFs by far is the estimation of the overall redshift distribution $N(z)$, the true value of which is known for the BUZZARD simulation and will be denoted as $N'(z)$. Though alternatives exist (Malz & Hogg prep), "stacking" according to

$$\hat{N}^H(z) \equiv \frac{1}{N_{tot}} \sum_{i}^{N_{tot}} \hat{p}_i^H(z) \tag{16}$$

is the most widely accepted method for obtaining the stacked estimator $\hat{N}^H(z)$ of the redshift distribution from photo-z PDFs derived by a method H. Though we do not endorse the use of the stacked estimator of the redshift distribution, we use it under the assumption that the response of our metrics of $\hat{N}^H(z)$ will be analogous to the same metrics applied to a principled estimator of the redshift distribution. We must note, however, that this is a poor assumption in general.

Returning to the prior of photo-z PDFs, the true redshift distribution satisfies the tautology $N'(z) = p(z|I_D)$, because our training data is representative, I_D represents a prior that is also equal to the truth. In this ideal case of representative training data, the method that would give the best approximation to $N'(z)$ would be one that neglects all the information contained in the photometry $\{d_i\}_{N_{tot}}$ and gives every galaxy the same photo-z PDF $\hat{p}_i(z) = N'(z)$ for all i. (In fact, including any information from the photometry would only add noise to the optimal result of returning the prior for every galaxy.) This is the exact estimator, TRAINZ, that we have described in Section 3.3, and which will serve as an experimental control.

The exact implementation of the stacked estimator $\hat{N}^H(z)$ should depend on the parametrization of the photo-z PDFs, which may differ across codes and can affect the precision of the estimator (Malz et al. 2018); even considering a single method under the same parametrization, say a piecewise constant function over bins or a set of samples from the posterior, an estimator using $2N$ bins or samples will trivially be more precise than an estimator using N bins or samples. In order to minimize the effects of the choice of parameterization, we asked those running all twelve codes to output photo-z PDFs parameterized with 200 piecewise constant bins spanning $0 < z < 2$. The piecewise constant format is chosen because of its established presence in the literature, and the choice of 200 bins was motivated by the approximate number of columns expected to be available for storage of photo-z PDFs for the final LSST Project tables.[17] All the photo-z PDF catalogs are processed using the `qp` software package (Malz et al. 2018)[18] for manipulating and calculating metrics of univariate PDFs. We will discuss the choice of photo-z PDF parameterization further in Section 5.

4.1 Metrics of an ensemble of photo-z posteriors

Because the photo-z PDFs in the LSST catalog will be used for many applications, some of which require accuracy of each individual catalog entry, we consider several metrics that get at the population-level performance of the photo-z PDFs as distinct from a summary statistic thereof.

4.1.1 Probability integral transform (PIT)

The probability integral transform (PIT) has been employed recently to evaluate fidelity of posterior distributions (e. g. Bordoloi et al. 2010; Polsterer et al. 2016; Tanaka et al. 2018). The PIT value is defined for each individual galaxy as:

$$\mathrm{PIT} = \int_{-\infty}^{z_{\mathrm{true}}} p(z)\, dz\,. \tag{17}$$

The distribution of PIT values quantifies the behavior of the *ensemble* of photo-z PDFs, enabling us to evaluate whether the population of photo-z PDFs is, on average, accurate. The PIT value for each galaxy is the Cumulative Distribution Function (CDF) of its photo-z PDF evaluated at its true redshift. A catalogue of photo-z PDFs that are accurate should have a flat PIT histogram (i.e., the individual PIT values as samples from each CDF should match a Uniform(0,1) distribution if the CDFs are accurate). Specific deviations from flatness indicate inaccuracy: overly broad photo-z PDFs would manifest as underrepresentation of the lowest and highest PIT values, whereas overly narrow photo-z PDFs would manifest as over-representation of the lowest and highest PIT values. High frequency at only PIT ≈ 0 and PIT ≈ 1 indicates the presence of catastrophic outliers with highly inaccurate photo-z PDFs where the true redshift of a galaxy is outside of the support of its photo-z PDF. Tanaka et al. (2018) use the histogram of PIT values as a diagnostic indicator of overall code performance, while Freeman et al. (2017) independently define the PIT and demonstrate how its individual values may be used both to perform hypothesis testing (via, e.g., the KS, CvM, and AD tests; see below) and to construct quantile-quantile plots.

4.1.2 Quantile-quantile (QQ) plot

A quantile is defined by partitioning a distribution into consecutive intervals containing equal amounts of probability, or equal numbers of objects in each interval in the case of a distribution of objects. The quantile-quantile (QQ) plot serves as a graphical visualization for comparing two distributions, where the quantiles of one distribution are plotted against the quantiles of the other distribution. The QQ plot provides an easy way to qualitatively assess the differences in various properties such as the moments of an estimating distribution relative to a true distribution.

In this paper, QQ plots are used for two purposes: (1) for comparing $N(z)$ from photo-z PDFs (estimated using Eq. 16) with the true $N(z)$, and (2) for assessing the overall consistency of an ensemble of photo-z PDFs with their true redshifts on a population level, where the distribution of the PIT values (see previous section) is compared to a uniform distribution between 0 and 1. Though the QQ plot contains very similar information to that shown in the PIT histogram plot, due to being the PIT being the derivative of the QQ, we include both forms for completeness and enhanced visual interpretability.

4.1.3 Conditional density estimation loss

With the conditional density estimation loss (CDE loss) we can compare how well different methods estimate individual PDFs for photometric covariates $\mathbf{x}$ rather than looking only at the ensemble distribution. As in Section 3.2.4, we use the notation $f(z|\mathbf{x})$ instead of $p(z)$ to explicitly show the dependence on the photometry $\mathbf{x}$.

The CDE loss is defined as

$$L(f, \widehat{f}) \equiv \int \int \left(f(z \mid \mathbf{x}) - \widehat{f}(z \mid \mathbf{x})\right)^2 dz dP(\mathbf{x}). \tag{18}$$

This loss is the CDE equivalent of the RMSE in regression. To estimate this loss we rewrite it as

$$L(f, \widehat{f}) = \mathbb{E}_{\mathbf{X}} \left[\int \widehat{f}(z \mid \mathbf{X})^2 dz\right] - 2\mathbb{E}_{\mathbf{X},Z} \left[\widehat{f}(Z \mid \mathbf{X})\right] + K_f, \tag{19}$$

where upper-case letters denote random variables and lower-case the observed variables. The first expectation is with respect to the marginal distribution of the covariates $\mathbf{X}$, the second expectation is with respect to the joint distribution of $\mathbf{X}$ and Z, and K_f is a constant depending only upon the true conditional densities $f(z \mid \mathbf{x})$. For each method we can estimate these expectations as empirical expectations on the test or validation data (Eq. 7 in Izbicki et al. 2017) without knowledge of the true densities.

4.2 Metrics over estimated probability distributions

In tandem with the QQ and PIT metrics introduced above, we additionally compute the following metrics comparing the empirical CDF of a distribution to the true or expected distribution. These metrics give a more quantitative measure of the departure from ideal than the more visual PIT histogram and QQ plot. We compute metrics comparing the CDF of PIT values to the CDF of a Uniform distribution, and also compute the CDF of the true redshift distribution $N'(z)$ compared the $\hat{N}(z)$ distribution derived from summing the photo-z PDFs as described in Eq. 16.

4.2.1 Kolmogorov-Smirnov (KS) and related statistics

The *Kolmogorov-Smirnov statistic* N_{KS} is the maximum difference between $F_{\mathrm{phot}}(z)$ and $F_{\mathrm{spec}}(z)$, the CDFs of the

[17] See, e. g. the LSST Data Products Definition Document, available at: `https://ls.st/dpdd`
[18] available at: `http://github.com/aimalz/qp/`

photo-z and spectroscopic redshift respectively:

$$N_{\mathrm{KS}} \equiv \max_{z} \left(|F_{\mathrm{phot}}(z) - F_{\mathrm{spec}}(z)| \right). \tag{20}$$

The KS test quantifies the similarity between two distributions, independent of binning. A lower N_{KS} value corresponds to more similar distributions.

We also consider two variants of the KS statistic: the Cramer-von Mises (CvM) and Anderson-Darling (AD) statistics. The CvM statistic is similar to the KS statistic as it is also computed from the distance between the measured CDF and the ideal CDF, but instead of the maximum distance, the CvM statistic

$$\omega^2 \equiv \int_{-\infty}^{+\infty} \left(F_{\mathrm{meas.}}(x) - F_{\mathrm{ideal}}(x) \right)^2 \mathrm{d}F_{\mathrm{ideal}} \tag{21}$$

is the average of the distance squared.

The AD statistic

$$A^2 \equiv N_{tot} \int_{-\infty}^{+\infty} \frac{\left(F_{\mathrm{meas.}}(x) - F_{\mathrm{ideal}}(x) \right)^2}{F_{\mathrm{ideal}}(x)(1 - F_{\mathrm{ideal}}(x))} \mathrm{d}F_{\mathrm{ideal}} \tag{22}$$

is a weighted version of the CvM statistic, making it more sensitive to the tails of the distribution, where N_{tot} is the sample size.

4.2.2 Moments

We additionaly calculate the first three moments of the estimated redshift distribution $\hat{N}^H(z)$ for each code and compare them to the moments of the true redshift distribution $N'(z)$. The m^{th} moment of a distribution is defined as

$$\langle z^m \rangle \equiv \int_{-\infty}^{\infty} z^m N(z) dz. \tag{23}$$

Here, we use the moments of the stacked estimator of the redshift distribution function as the basis for a metric. The closer the moments of $\hat{N}(z)$ for a photo-z PDF method are to the moments of the true redshift distribution function $N'(z)$, the better the photo-z PDF method is a estimating the overall redshift distribution "shape".

5 RESULTS

5.1 Ensembles of photo-z interim posteriors

Fig. 1 Shows the $p(z)$ produced by each of our twelve photo-z codes for four example galaxies which exemplify some prominent cases that arise when estimating photo-z PDFs: a narrow, unimodal redshift solution, a broader unimodal solution, a bimodal distribution, and a complex, multimodal distribution. The red vertical line represents the true redshift of the individual galaxy, and the blue curve represents the redshift probability. Several features are obvious even in these illustrative examples. ANNz2, METAPhoR, CMNN, and SkyNet all show an excess of small-scale features, which appear to be print-through of the underlying training set galaxies. For example, in CMNN the $p(z)$ are a simply a weighted histogram of all spectroscopic training galaxies in nearby colour space with no smoothing applied, so the substructure is due to the finite number of neighbours, and is not unexpected. GPZ (in its current implementation), on the other hand, always produces a single Gaussian, which broadens to cover the multi-modal redshift solutions seen in other codes.

As stated in Section 4, $p(z)$ is parameterized as 200 piecewise constant bins covering $0 < z < 2$ for all twelve codes, giving a grid size of $\delta z = 0.01$ for each PDF. A piecewise constant grid was a natural choice for some photo-z methods, for instance most template-based codes compute likelihoods on a fixed grid. For faint galaxies, this redshift resolution is sufficient to capture the shape of $p(z)$ for the majority of the test sample, where photometric errors on the faint galaxies lead to somewhat broad peaks in the redshift posterior. However, as can be seen in e. g. the top left panel of Fig. 1, for bright galaxies with narrow $p(z)$ the grid spacing of $\delta z = 0.01$ is not always sufficient to resolve the peak. This is consistent with the results described in Malz et al. (2018), who find that quantiles (and, to a lesser degree, samples) often outperform gridded $p(z)$, particularly for bright objects and in the presence of harsher storage constraints. With a full 200 numbers to capture the information of each photo-z PDF, any parametrization will perform adequately, but other storage parametrizations and limits on storage resources may be considered in future work. We will discuss this further in Section 6.

Fig. 2 shows both the quantile-quantile plots (red) and the histogram of PIT values (blue) summarizing the results from each photo-z code. The red line shows the measured quantiles, while the black diagonal represents the ideal QQ values if the distribution were perfectly reproduced. A second panel below the main panel for each code shows the difference between Q_{data} and Q_{theory}, i. e. the departure from the diagonal, for clarity. Biases and trends in whether the average width of the $p(z)$ values being over/under-predicted are evident. An overall bias where the predicted redshift is systematically low manifests as the measured QQ value falling above the diagonal, as is the case for BPZ and EAZY, while a systematic overprediction shows up as the measured QQ value falling below the diagonal, as seen in TPZ. In terms of PIT histograms, a systematic underprediction of redshift corresponds to fewer PIT values at $PIT < 0.5$ and more at $PIT > 0.5$, while a systematic overprediction will show the opposite.

Examination of the PIT histograms and QQ plots shows that there are fairly generic issues with the width of $p(z)$ uncertainties: Delight, CMNN, SkyNet and TPZ all show a PIT histogram with a dearth of low values and and an excess of high values, signs that, on average, their $p(z)$ are more broad than the true distribution of redshifts. METAPhoR shows the opposite trend, indicating the the $p(z)$ are more narrow than the distributions given by the true redshifts. In all of these code cases there is a free parameter or bandwidth that can be used to tune uncertainties. The sensitivity of multiple codes to this bandwidth choice emphasizes the fact that great care must be taken in setting user-defined parameters in photo-z codes, even in the presence of representative training/validation data. for FlexZBoost the "sharpening" parameter (described in Section 3.2.4) plays a key role in improving the results, resulting in a QQ plot that is very nearly diagonal. A similar sharpening procedure could be beneficial for several codes. Interestingly, the three purely template-based codes, BPZ, EAZY, and LePhare, show relatively well behaved $p(z)$ statistics (albeit with some bias), which may indicate that the likelihood estimation with

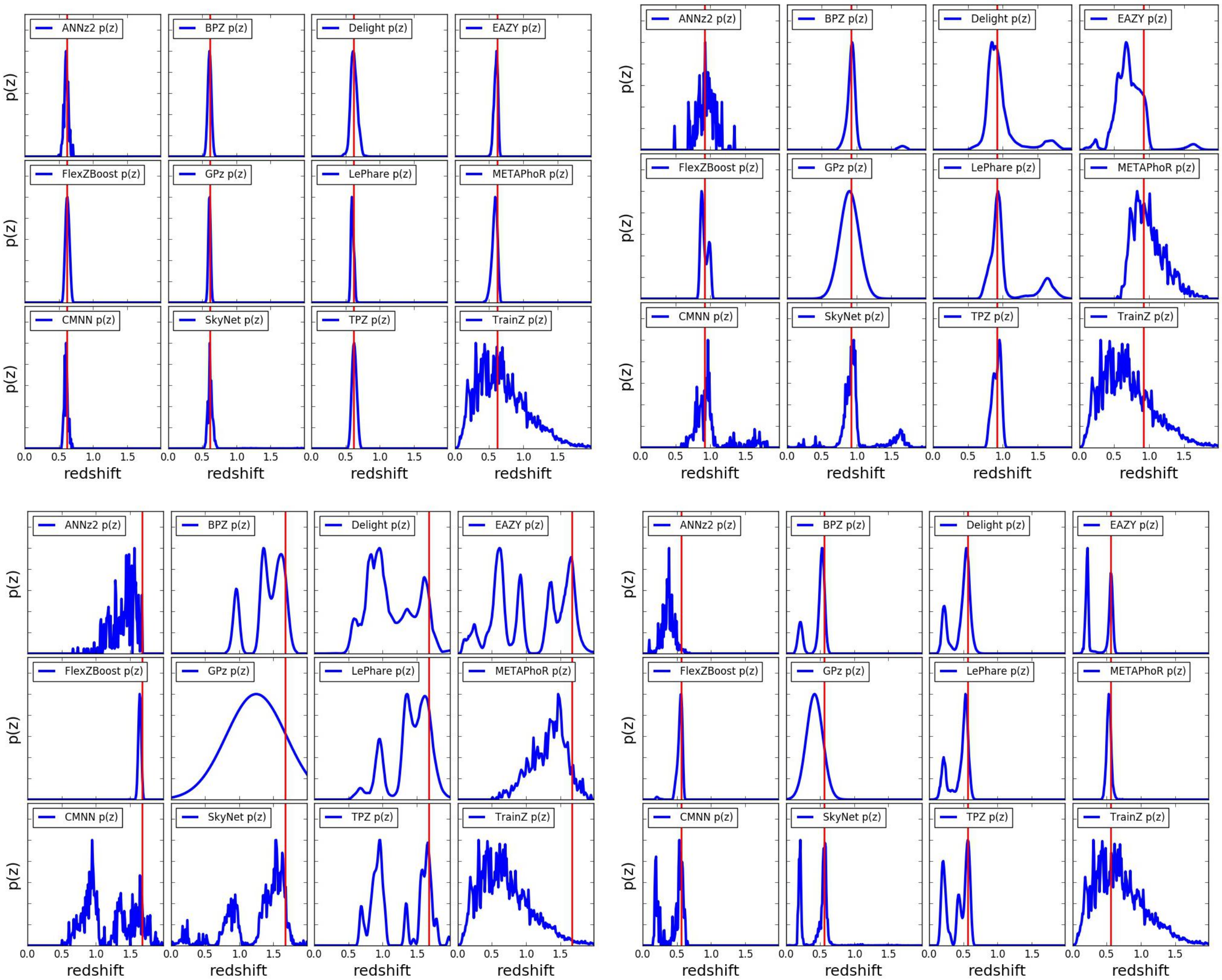

Figure 1. Four illustrative examples of individual p(z) distributions produced by the codes. The red vertical line represents the true redshift. Examples are chosen with common features seen in PDFs: tight unimodal $p(z)$ (upper left), broad unimodal $p(z)$ (upper right), bimodal $p(z)$ (lower right), and complex/multimodal $p(z)$ (lower left). Codes show varying amounts of small-scale structure in their reconstruction of the posterior distribution. We see varying responses from the codes in the presence of color degeneracies and photometric errors, resulting in narrow and broad unimodal, bimodal, and multi-modal $p(z)$ curves.

representative templates is accurately capturing the uncertainties on individual redshifts.

The ideal PIT histogram would follow the black dashed line, representing a uniform distribution of PIT values, equivalent to the diagonal line in the QQ plot. Overly broad $p(z)$ values show up as an excess of PIT values near 0.5 and a dearth of values at the edges, while overly narrow $p(z)$ will have an excess at the edges and will be missing values at the centre. Another feature evident in the PIT histograms is the number of "catastrophic outlier" values where the true redshift falls outside of the non-zero support of $p(z)$, corresponding to PIT= 0.0 or 1.0 is more apparent than in the QQ plots. Following Kodra & Newman (in prep.) we define f_{O} as the fraction of objects with $PIT < 0.0001$ or $PIT > 0.9999$. Table 2 lists these fractions for each of the codes. For a proper Uniform distribution we expect a value of 0.0002. Several codes show a marked excess, with ANNz2, FlexZBoost, LePhare, and METAPhoR with $f_{\mathrm{O}} > 0.02$, indicating a sizeable number of catastrophic redshift solutions where the true redshift is not covered by the extent of $p(z)$. For METAPhoR this may be partially due to an overall underprediction of the $p(z)$ width, however this is not the case for the other codes. LePhare is a particular outlier with nearly 5 per cent of objects outside of $p(z)$ support. Further study will be necessary to determine what is causing these misclassifications for LePhare. As expected, and by design, TrainZ has the proper fraction of outliers for the f_{O} statistic.

Fig. 3 shows comparative metric values for the quantitative Kolmogorov-Smirnoff (KS), Cramer-Von Mises (CvM), and Anderson Darling (AD) test statistics for each of the codes based on comparing the distribution of their PIT values to the expected uniform distribution over the interval [0,1]. The individual values of the statistic are not as important as the comparative score between the different codes. The AD test statistic diverges for values that include the ex-

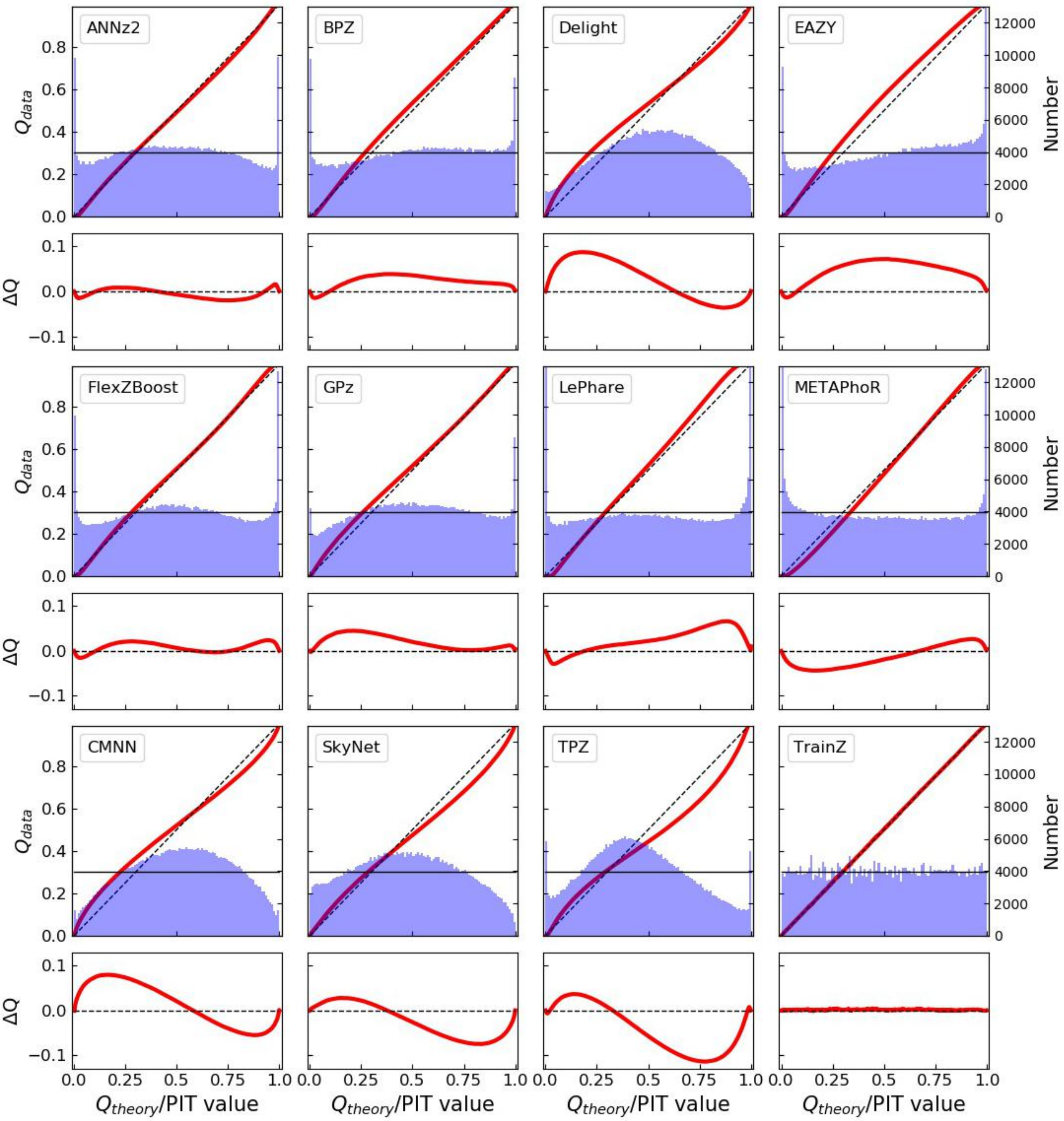

Figure 2. Summary plots for all twelve photo-z codes illustrating performance for the interim posterior statistics. The top panel of each pair shows both the Quantile-Quantile (QQ) plot (red) and the histogram of PIT values (blue). The desired behavior is a QQ plot that matches the diagonal dashed line, and a PIT histogram that matches a uniform distribution matching the thin horizontal black line. The bottom panel of each pair shows the difference between the QQ quantile and the diagonal, illustrating departure from the desired performance. Histograms with an overabundance of PIT values at the centre of the distribution indicate $p(z)$ distributions that are overly broad, while an excess of values at the extrema indicate $p(z)$ distributions that are overly narrow. Values of PIT=0 and PIT=1 indicate "catastrophic failures" where the true redshift is completely outside the support of $p(z)$. Asymmetric features are indicative of systematic bias in the redshift predictions. A variety of behaviors are evident, and specific details are discussed in the text.

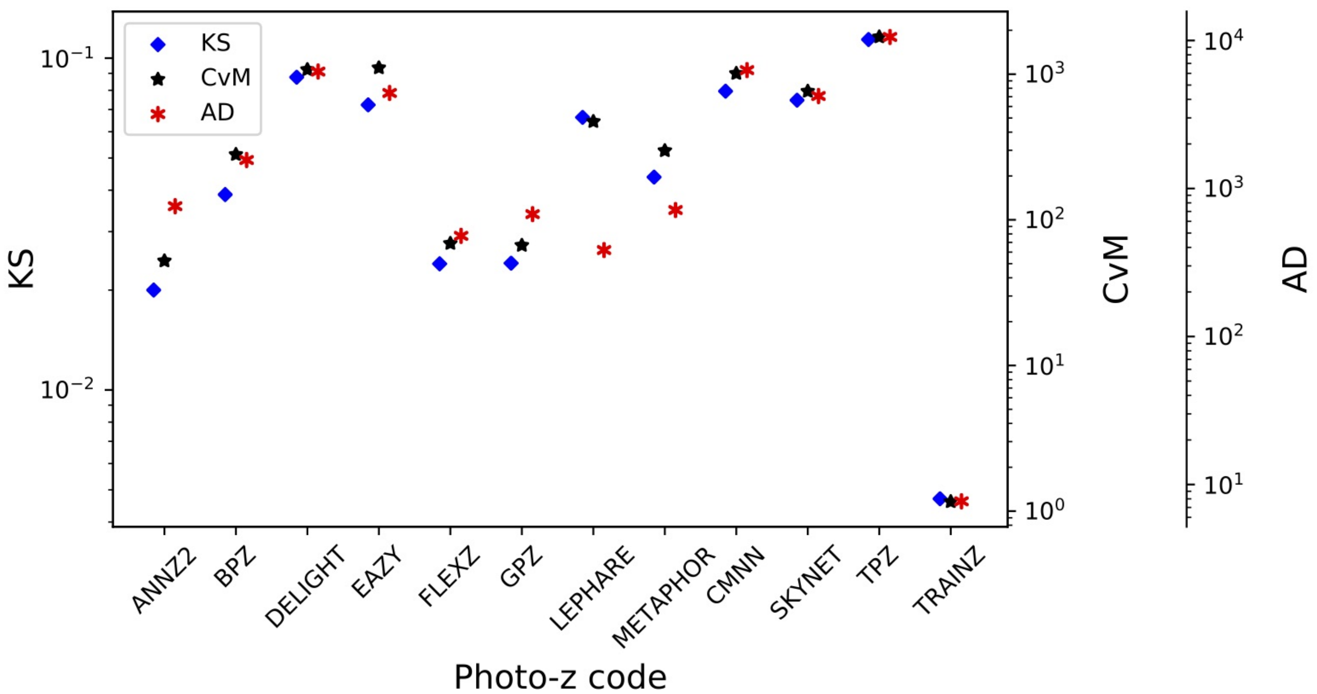

Figure 3. A visual representation of the Kolmogorov-Smirnoff (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the PIT distributions. The statistics are often highly correlated, with similar relative values between metrics for each code; however, the AD statistic excludes some values at the extrema of the distributions, and can have disparate values compared to KS and CvM.

Table 2. The fraction of "catastrophic outlier" PIT values. We expect a value of 0.0002 for a proper Uniform distribution. An excess over this small value indicates true redshifts that fall outside the non-zero support of the $p(z)$.

"catastrophic outlier" PIT fraction	
Photo-z Code	fraction PIT$< 10^{-4}$ or >0.9999
ANNz2	0.0265
BPZ	0.0192
Delight	0.0006
EAZY	0.0154
FlexZBoost	0.0202
GPz	0.0058
LePhare	0.0486
METAPhoR	0.0229
CMNN	0.0034
Skynet	0.0001
TPZ	0.0130
TrainZ	0.0002

trema, and thus is calculated by excluding the edges of the distribution. We calculate the AD statistic over the range of PIT values $v = [0.01, 0.99]$. ANNz2 and FlexZBoost score very well for the PIT metrics. METAPhoR and LePhare score very well in the PIT AD statistic, but both have a large number of catastrophic outliers, resulting in higher KS and CvM scores.

Given the near-perfect training data, examining the individual codes for explanations for departures from the expected behaviour will be instructive in avoiding similar problems in future tests. ANNz2 performs quite well in $p(z)$ based metrics. In the specific implementation employed in this paper, the final $p(z)$ is a weighted average of five neural-nets. During the training process ANNz2 compares the percentiles of the redshift training sample against the CDFs of the $p(z)$ sample. Distributions that more closely match are given extra weight, and the final weights are designed to produce accurate percentiles. Given that our metrics are focused on the percentile distributions, it is unsurprising that ANNz2 performs well in the given metrics. The discreteness in the individual $p(z)$ estimated by ANNz2 can be attributed to the fact that the code was run as a classifier, assigning weights to discrete bins of redshift. While multiple bins may receive weight, the bins themselves will still be discretized, and no additional smoothing was performed. Overall, FlexZBoost and ANNz2 show the best ensemble agreement in their distribution of PIT values.

5.2 Metrics of the stacked estimator of the redshift distribution

Fig. 4 shows the stacked $\hat{N}(z)$ distribution compared to the true redshift distribution $N'(z)$ for all tested codes. The red line indicates the summed $p(z)$ for each code, while the blue line shows the true redshift distribution. All distributions are smoothed via kernel density estimation (KDE) with a common bandwidth chosen via Scott's rule (Scott 1992) in order to minimize differences in small-scale features and make for

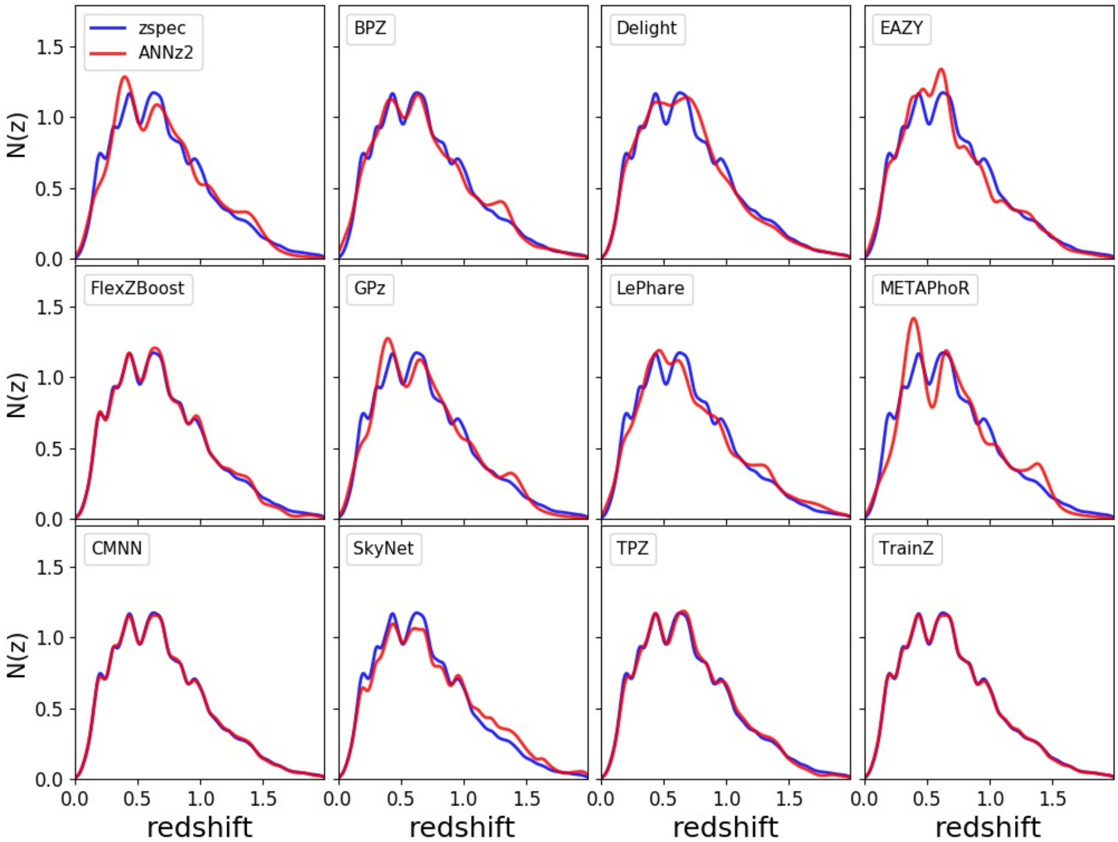

Figure 4. The stacked $p(z)$ produced by each photo-z code ($\hat{N}(z)$, red) compared to the spectroscopic redshift distribution ($N'(z)$, blue). Varying levels of agreement are seen in the codes. Both $\hat{N}(z)$ and $N'(z)$ in all codes are smoothed using a single bandwidth chosen via Scott's rule.

a more uniform comparison between codes. While Scott's rule is used to display $N'(z)$ in the figure, all quantitative statistics are computed via the empirical CDF, and are thus unaffected by bandwidth/smoothing choice. As expected, TrainZ is in excellent agreement with the true redshift distribution: as the training sample is selected from the same underlying distribution as the test set, the redshift distributions are identical, up to Poisson fluctuations due to the finite number of sample galaxies. CMNN is also in excellent agreement for similar reasons: with a representative training sample of galaxies spanning the colour-space, the sum of the colour-matched neighbour redshifts should return the true redshift distribution. FlexZBoost and TPZ also show very good agreement, with only slight departures, with an over/under-prediction in the high redshift tail of $\hat{N}(z)$ evident around $z \sim 1.4$. In fact, several of the other codes show an excess at $z \sim 1.4$, particularly the template-based codes BPZ, EAZY, and LePhare. This is likely due to the 4000 Å break passing through the gap between the z and y filters, resulting in a drastic change in $z-y$ colour for galaxies in this redshift range. With a relative dearth of strong features blue-ward of the 4000 Å break in most galaxy SEDs, the colour change in the two reddest filter bandpasses of a survey has a large influence on the redshift determination. The $z \sim 1.4$ feature is one of the most prominent sources of larger uncertainty in individual galaxy $p(z)$. In our sample individual galaxy $p(z)$'s tend to be broader around $z \sim 1.4$ and point estimates are more uncertain in this regime, as is readily seen in the point-estimate plots shown in Fig. A1 and described in the Appendix. This feature is not unique to this dataset, it is a common occurrence in photo-z estimation. The fact that similar excesses appear in Figure 4 for ANNz2 and METAPhoR shows that the effect is not limited to template-based codes. However, the lack of such a feature in the other codes shows that it is possible to eliminate the degeneracies. Further study on this issue may provide a solution for codes that suffer from this shortcoming.

Two of the machine learning based codes, ANNz2 and METAPhoR, appear to be over-trained, adding excess galaxy probability to the redshift peaks with the largest number of training galaxies, and missing probability in the troughs where training galaxies are of fewer number. Given

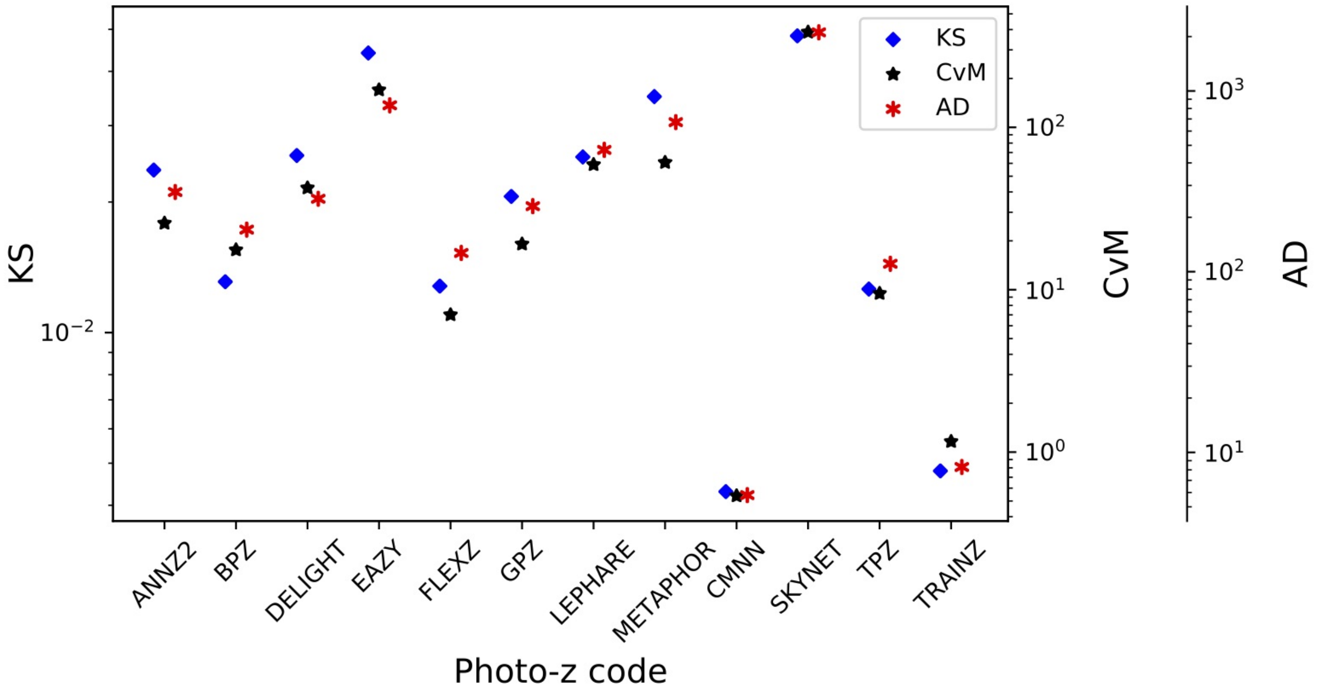

Figure 5. A visual representation of the Kolmogorov-Smirnoff (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the $\hat{N}(z)$ distributions. The statistics are correlated, the codes with the lowest KS statistics tend to have the lowest CvM and AD statistics. CMNN performs markedly better than the others in reconstructing the overall $N(z)$ distribution, while SkyNet scores poorly due to an overall bias in its redshift predictions.

that our training data is drawn from the same galaxy population as the test set, and our data has prominent peaks in $N'(z)$, perhaps it is not unexpected that such overtraining occurs in some codes, though the fact that it does not occur in all training-based codes indicates that it may be due to specifics of the implementations and bandwidth choices of ANNz2 and METAPhoR. This, once again, emphasizes that care much be taken in choice of bandwith parameters in individual codes in order to avoid such overtraining. SkyNet shows an obvious redshift bias, evident both visually in Figure 4 and in the first moment of N(z) listed in Table 4, where it is clearly an outlier. SkyNet employed a method where a random sample of training galaxies was chosen, but there was no test that the subset was completely representative of the overall redshift distribution. Unlike the previous implementation of SkyNet in Bonnett (2015), no effort was made to add extra weight to more rare low and high redshift galaxies. Either of these decisions could be the cause of the bias seen in our results. Future runs of SkyNet will explore these implementation choices and their effects.

Figure 5 shows the quantitative Kolmogorov-Smirnoff (KS), Cramer-Von Mises (CvM), and Anderson Darling (AD) test statistics for each of the codes for the $\hat{N}(z)$ based measures. FlexZBoost, CMNN, and TPZ outperform the other codes in the $\hat{N}(z)$ metrics. It is unsurprising that CMNN scores well, as with a near perfectly representative training set means that choosing neighbouring points in color/magnitude space should lead to excellent agreement in the final $\hat{N}(z)$ estimate. TPZ performed quite poorly in $p(z)$ statistics, but results in a good fit to the overall $N(z)$. This is somewhat surprising, as performance was optimized for accurate $p(z)$, not $\hat{N}(z)$. During the validation stage for TPZ, there was a trade off between the width of the $p(z)$ when adjusting a smoothing parameter and overall redshift bias. The optimal result in the PIT metrics, as illustrated in the shape of the QQ plot, does contain some level of bias as well as a slight underprediction of mean $p(z)$ width, which translates to poor metric scores. This is something that will be looked into for TPZ in the future.

Table 3 shows the CDE loss statistic for each photo-z code. Once again FlexZBoost and CMNN score very well for the stacked $\hat{N}(z)$ metrics, as do GPz and TPZ. The CDE loss measures how well individual PDFs are estimated, and codes with a low CDE loss tend to have good $\hat{N}(z)$ estimates (though the reverse is not necessarily true). FlexZBoost is optimized to minimize CDE loss which may explain why the method has good ensemble metrics as well. Note from Table 3 that both FlexZBoost and CMNN have low CDE losses. As CDE loss is a better measure of individual redshift performance, rather than ensemble distribution performance, this statistic is a better indicator of which codes will be most likely to perform well for science cases where single objects are employed.

Table 4 lists the first three moments of the stacked $\hat{N}(z)$ distribution, including the moments of the "truth" distribution for comparison. Several codes are able to reproduce the mean and variance of the distribution to less than a per cent, while several codes do not, which may be a cause for concern, given that mean and variance of the redshift distribution are key properties in cosmological analyses. We note that this stated goal of the study as defined for participants was to accurately reproduce $p(z)$, the "stacking" of the probability distributions to estimate $\hat{N}(z)$ was not the focus as stated to the participants. This explains why some of the best-performing empirical codes in terms of $p(z)$ measures (e. g. FLEXZBOOST) do not do as well at reproducing $\hat{N}(z)$ moments. Had we defined a different parameter to optimize, in this case overall accuracy of $\hat{N}(z)$ rather than individual $p(z)$, would result in improved performance in a particular metric. That is, optimizing photo-z performance for one metric does not automatically give optimal performance for other metrics. As previously stated, there are a variety of scientific use cases for photo-z's in large upcoming surveys, and care must be taken in how the metrics used to optimize catalog photometric redshifts are defined as well as in how they are used. This implies that we may need multiple photo-z estimators, tuned to the particular metric, in order to maximize returns over all science cases in upcoming surveys. In addition, very few scientific use cases will employ the overall $\hat{N}(z)$ with no cuts, as we explore in this paper. We discuss more realistic tomographic bin selections that will be explored in a follow-up paper in Section 6.1.

5.3 Interpretation of metrics

Samples from accurate photo-z posteriors should reproduce the space of $p(z, \text{data})$. However, it is difficult to test this reconstruction given our data set, as the galaxy distributions arise from mock objects pasted on to an underlying dark matter halo catalogue with properties designed to match empirical relations, rather than being drawn from statistical distributions in redshift. In previous sections we have mentioned that optimizing for a specific metric does not guarantee good performance on other metrics, nor is there any guarantee that good performance by our metrics corresponds to *accurate* photo-z posteriors, in the sense of predicting redshifts for individual galaxies. In other words, we can construct photo-z estimators that provide good coverage in many of our tests, but which have very little predictive power.

The TRAINZ estimator, which assigns every galaxy a $p(z)$ equal to $N(z)$ of the training set as described in Section 3.3, is introduced as a "null test" to demonstrate this point via *reductio ad absurdum*. TRAINZ outperforms all codes on the PIT-based metrics, and all but one code on the $N(z)$ based statistics. Because our training set is perfectly representative of the test set, $N(z)$ should be identical for both sets down to statistical noise.

The CDE loss and point estimate metrics, however, successfully identify problems with TRAINZ. As shown in Appendix A, TRAINZ has identical $ZPEAK$ and $ZWEIGHT$ values for every galaxy, and thus the photo-zs are constant as a function of spec-zs, i.e. a horizontal line at the mode and mean of the training set distribution respectively. The explicit dependence on the *individual* posteriors in the calculation of the CDE loss, described in Section 4.1.3, distinguishes this metric from the other $p(z)$ metrics that test the overall ensemble of $p(z)$ distributions. With a representative training set, TRAINZ will score well on the ensemble metrics, but fails miserably for metrics tied to individual redshifts presented in this paper. We note that many of the ensemble-based metrics are prominent in the photo-z literature despite their inability to identify problems such as those exemplified by TRAINZ.

In summary, context is crucial to interpreting metrics and defending against deceptively well performing methods such as TRAINZ. The best photo-z method is the one that most effectively achieves our science goals, not the one that performs best on a metric that does not accurately reflect those goals. In the absence of clear goals or the information necessary for a principled metric definition, we must think carefully before choosing a single metric.

Table 3. CDE loss statistic for each photo-z code.

Photo-z Code	**CDE Loss**
ANNz2	−6.88
BPZ	−7.82
DELIGHT	−8.33
EAZY	−7.07
FLEXZBOOST	−10.60
GPz	−9.93
LEPHARE	−1.66
METAPHoR	−6.28
CMNN	−10.43
SKYNET	−7.89
TPZ	−9.55
TRAINZ	−0.83

Table 4. Moments of the stacked $\hat{N}(z)$ distribution

Stacked $n(z)$ Moments			
	1st Moment	2nd Moment	3rd Moment
TRUTH	0.701	0.630	0.671
Photo-z Code	1st Moment	2nd Moment	3rd Moment
ANNz2	0.702	0.625	0.653
BPZ	0.699	0.629	0.671
DELIGHT	0.692	0.609	0.638
EAZY	0.681	0.595	0.619
FLEXZBOOST	0.694	0.610	0.631
GPz	0.696	0.615	0.639
LEPHARE	0.718	0.668	0.741
METAPHoR	0.705	0.628	0.657
CMNN	0.701	0.628	0.667
SKYNET	0.743	0.708	0.797
TPZ	0.700	0.619	0.643
TRAINZ	0.699	0.627	0.666

6 DISCUSSION AND FUTURE WORK

In this paper we presented results evaluating the computation of individual galaxy photometric redshift PDFs for twelve photo-z codes. As discussed in Section 4 each $p(z)$ should accurately reflect the relative likelihood as a function of redshift for each galaxy in an informative way; that is, the estimates should provide useful information per individual galaxy, not just the ensemble. All codes were provided a set of representative training data and tested on an idealized set of model galaxies with high signal-to-noise and photometry with no confounding effects due to blending, instrumental effects, the night sky, or other complications included. In many ways, this was a baseline test of a "best case scenario" to predict the expected photo-z performance if a stage IV dark energy survey was to obtain complete training samples and perfectly calibrated their multi-band photometry. Given these idealized conditions, any deficiencies observed in a photo-z code's performance should be a cause for concern, and may be evidence of a problem with either/both of the specific code implementation or the underlying algorithm. Many of the codes tested performed well; however, several did not meet the stringent goals that have been laid out for LSST photometric redshift performance for individual galaxies, as laid out in the LSST SRD (See Section 1).If methods can not reach the goals on idealized data, then they will almost surely not meet those same goals when the more complex problems that we expect to arise from real LSST data are included. The results presented in this paper enable an evaluation of which algorithms are the most promising moving forward, and potentially point to implementation choices or mistakes which could be improved or corrected in others. Codes that do not perform well may not be worth pursuing in future challenges.

One obvious trend in several of the codes tested was an overall over or underprediction of the widths of $p(z)$, as evidenced by the QQ plots and PIT histograms shown in Fig. 2. A more careful tuning of bandwidth or smoothing during the validation process appears to be necessary for many of the machine learning based codes in order to improve the accuracy of $p(z)$. For narrow peaked $p(z)$ the parameterization of the PDF as evaluated on a fixed redshift grid could also have contributed to some overestimates of $p(z)$ width simply due to the finite resolution. After evaluating results such as those presented in Malz et al. (2018), in future analyses we plan to switch from a fixed grid to quantile-based storage of $p(z)$ in order to more efficiently and accurately store redshift PDF results.

Another important factor to keep in mind when examining the results presented in this paper is the fact that they are at some level dependent on the metrics that we aim to optimize: in this case code participants were asked to submit their optimal measures of an accurate $p(z)$, so participants used the training/validation data to optimize their codes accordingly. Had we, instead, asked for an optimal $\hat{N(z)}$ the resulting metrics would be different for most, if not all, of the codes, as they would optimize toward a different goal. Specific metric choice can affect which codes are among the "best" codes. As stated earlier, there are cosmological science cases that require either individual galaxy photo-z measures, or ensemble $\hat{N}(z)$ measures. We must be aware of that the optimal method for one is not necessarily optimal for the other, and in fact several photo-z algorithms may be necessary in the final cosmological analysis in order to satisfy the requirements of all science use cases.

The example of the simple TRAINZ estimator described in Section 5.3 shows a simple model with a $p(z)$ that is unrealistic for individual objects can still score very well on many of our metrics. It is important to look at *all* metrics, and keep in mind what information each metric conveys. We re-emphasize that the dataset tested was quite idealized, and discuss enhancements that will be added in future simulations to test photo-z codes on increasingly realistic conditions in the following section.

6.1 Future work

The work presented in this paper is only the first step in characterizing current photo-z codes and moving toward an improved photometric redshift estimator. This initial paper explored code performance in idealized conditions with perfect catalog-based photometry and representative training data. As mentioned in Section 5.2 for the stacked $N(z)$ metrics we examined only the entire galaxy population with no selections in either photo-z "quality" or redshift. The cosmological analyses for weak lensing and large scale structure based measures plan to break galaxy samples into tomographic redshift bins, using photo-z $p(z)$ to infer the redshift distribution for each bin. The specific selection used to determine these bins, both algorithmically and the specific bin boundaries, could induce biases due to indirect selections inherent in the photo-z or other bin selection parameters. The effects of tomographic bin selection will be explored in a dedicated future paper, including propagation of redshift uncertainties in a set of fiducial tomographic redshift bins in order to estimate impact on cosmological parameter estimation.

In future papers a focus of the *LSST Dark Energy Science Collaboration Photo-z Working Group* will be to add more and more complexity to our simulated data in order to test photo-z algorithms in increasingly realistic conditions. The most pressing concern is the impact of incomplete spectroscopic training samples. The SEDs for the galaxy sample in this paper were constructed form linear combinations of five basis SED templates. Future simulations will also include more complex SED information, with a more realistic range of physical properties, and the inclusion of AGN effects, a more insidious problem, where AGN features may not be apparent, but the colors and other host galaxy properties are perturbed relative to galaxies with an inactive nucleus. In such cases, the presence of the AGN may induce a bias if the template SEDs or empirical datasets do not include low-level AGN counterparts.

As discussed extensively in Newman et al. (2015) a representative set of spectroscopically confirmed galaxies spanning the full range of both redshift and apparent magnitude is necessary as a training set to characterize the mapping from broad-band fluxes to photometric redshifts. Current and upcoming surveys are putting significant effort into obtaining these training samples (e. g. Masters et al. 2017), however we still expect significant incompleteness for LSST-like samples, particularly at faint magnitudes. We plan to produce a realistically incomplete training set of spectroscopic galaxies, modeling the performance of spectrographs, emission-line properties, and expected signal-to-noise to determine which galaxies will fail to yield a secure redshift. In addition to outright redshift failures we will model the inclusion of a small number of falsely identified secure redshifts where misidentified emission lines or noise spikes cause an incorrect redshift solution to be marked as a high quality identification. Even sub-per cent level contamination by false redshifts can impact photo-z solutions at levels comparable to the stringent requirements of some LSST science cases. We expect different systematics to occur in different photo-z codes in response to training on incomplete data, particularly some of the machine learning methods. The response of the codes will inform future directions of code development.

The underlying dataset limited this work to a maximum redshift of $z = 2$. LSST imaging after 10 years of observations will include a significant number of $z > 2$ galaxies in expected cosmology samples, and their inclusion does have potential significant implications for photo-z measures: the high redshift galaxies lie at fainter apparent magnitudes and can have anomalous colours due to evolution of stellar populations and the shift to rest-frame magnitudes probing UV features of the underlying SED. More importantly, one of the most common "catastrophic outlier" degeneracies observed in deep photometric samples occurs when the Lyman break is mistaken for the Balmer break, leading to multiple redshift solutions at $z \sim 0.2 - 0.3$ and $z \sim 2 - 3$ (Massarotti et al. 2001). This degeneracy, along with other potential degeneracies, are currently not covered by the limited redshift range of this initial paper, which could mean that we are not probing the full range of potential extreme outlier populations and how our photo-z estimators respond to them. Extending simulations to include the high-redshift galaxy population will be a priority in future data challenges.

This initial paper explored a data set that was constructed at the catalog level, with no inclusion of the complications that come from measuring photometry from images. Future data challenges will move to catalogs constructed from mock images, including effects that will have great impact on photo-z measurements, which will naturally include the complications of object blending, sensor effects, different observing conditions, amongst others. Object blending will be a major area of investigation, as the mixing of flux from multiple objects and the resultant change in measured colours is predicted to affect a large fraction of LSST galaxies (Dawson et al. 2016), and will be one of the major contributing systematics for photo-z's.

Finally, while this paper and future papers discussed above focus on photometric redshift codes and estimating accurate $p(z)$ from training data, we plan a separate, but complementary, project to examine calibration of the resultant redshifts via spatial cross-correlations (Newman 2008), which will be explored in a separate set of papers. The overarching plan describing everything laid out in this section is described in more detail in the LSST DESC Science Roadmap (see Footnote in Section 1). These plans will require significant effort, but they are necessary if we are to make optimal use of the LSST data for astrophysical and cosmological analyses.

Acknowledgments

Author contributions are listed below.

S.J. Schmidt: Led the project. (conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, visualization, writing – original draft, writing – review & editing)

A.I. Malz: Contributed to choice of metrics, implementation in code, and writing. (conceptualization, methodology, project administration, resources, software, visualization, writing – original draft, writing – review & editing)

J.Y.H. Soo: Ran ANNz2 and Delight, updated abstract, edited sections 1 through 6, added tables in Methods and Results, updated references.bib and added references throughout the paper

M. Brescia: main ideator of METAPHOR and of MLPQNA; modification of METAPHOR pipeline to fit the LSST data structure and requirements

S. Cavuoti: Contributed to choice and test of metrics, ran METAPHOR, minor text editing

G. Longo: Scientific advise, test and validation of the modified METAPHOR pipeline, text of the METAPHOR section

I.A. Almosallam: vetted the early versions of the data set and ran many photo-z codes on it, applied GPz to the final version and wrote the GPz subsection

M.L. Graham: Ran the colour-matched nearest-neighbours photo-z code on the Buzzard catalog and wrote the relevant piece of Section 2; participated in discussions of the analysis.

A.J. Connolly: Developed the colour-matched nearest-neighbours photo-z code; participated in discussions of the analysis.

E. Nourbakhsh: Ran and optimized TPZ code on the Buzzard catalog and wrote a subsection of Section 2 for that

J. Cohen-Tanugi: contributed to running code, analysis discussion, and editing,reviewing the paper

H. Tranin: contributed to providing SkyNet results and writing the relevant section

P.E. Freeman: Contributed to choice of CDE metrics and to implementation of FlexZBoost

K. Iyer: assisted in writing metric functions used to evaluate codes

J.B. Kalmbach: Worked on preparing the figures for the paper.

E. Kovacs: Ran simulations, discussed data format and properties for SEDs, dust, and ELG corrrections

A.B. Lee: Co-developed FlexZBoost and the CDE loss statistic, wrote text on the work, and supervised the development of FlexZBoost software packages

C. Morrison: Managerial support; Discussions with authors regarding metrics and style; Some coding contribution to metric computation.

J. Newman: Contributions to overall strategy, design of metrics, and supervision of work done by Rongpu Zhou

E. Nuss: contributed to running code, analysis discussion, and editing,reviewing the paper

T. Pospisil: Co-developed FlexZBoost software and CDE loss calculation code

M.J. Jarvis: Contributed text on AGN to Discussion section and portions of GPz work

R. Izbicki: Co-developed FlexZBoost and the CDE loss statistic, and wrote software for FlexZBoost

The authors would like to thank their LSST-DESC publication review committee for comments that improved the paper draft.

personal funding sources S. Schmidt acknowledges support from DOE grant DE-SC0009999 and NSF/AURA grant N56981C. AIM is advised by David W. Hogg and was supported by National Science Foundation grant AST-1517237.

In addition to packages cited in the text, analyses performed in this paper used the following software packages: NUMPY and SCIPY (Oliphant 2007), MATPLOTLIB (Hunter

2007), SEABORN (Waskom et al. 2017), MINFUNC (Schmidt 2005), PYSKYNET (Bonnett 2016), and PhotUtils from the LSST simulations package (Connolly et al. 2014).

The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique des Particules in France; the Science & Technology Facilities Council in the United Kingdom; and the Department of Energy, the National Science Foundation, and the LSST Corporation in the United States. DESC uses resources of the IN2P3 Computing Center (CC-IN2P3–Lyon/Villeurbanne - France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515.

APPENDIX A: POINT ESTIMATE PHOTOMETRIC REDSHIFTS

While we assume that all science analysis will use full PDF information and do not recommend the use of single point estimates of redshift for most science applications, we include a brief evaulation as an Appendix. Plots of the point estimates can be a useful qualitative diagnostic of photo-z code performance, i. e. examining point photo-z vs. spec-z plots visually can give a quick impression of some common trends in different codes. Computing point estimate statistics may also be useful for more direct historical comparisons from older photo-z evaluations. If a point-estimate is preferred for a specific science case, it is fairly simple to compute the mean, mode, or some other simple estimator from each $p(z)$, so these point estimates can be easily derived from the stored $p(z)$.

There are several common point estimators of photo-z posteriors employed by different codes, e. g. the mode, mean, median of the $p(z)$ distribution. In addition, many of the machine learning based estimators can be set up to return a single redshift solution. For example, SkyNet can be configured to run as a regressor that returns a single float rather than a classifier that returns a 200-bin $p(z)$ estimate. The single value returned by a machine learning based code may not correspond to a particular measure such as the mode or mean, and so to avoid interpretation of results that might be introduced by variations in choice of specific point-estimate implementation per code, we discard the code-specific point estimates. We instead calculate point estimates more uniformly across the codes directly from the $p(z)$ using two measures, z_{PEAK} and z_{WEIGHT}. z_{PEAK} is simply the maximum value attained for each galaxy p(z), the mode of the probability distribution. z_{WEIGHT} is defined similarly to how it is defined in Dahlen et al. (2013), as the weighted mean of the redshift over the *main peak* of $p(z)$ containing the z_{PEAK} value. The main peak is defined by subtracting $0.05 \times z_{PEAK}$ from $p(z)$ and identifying the roots to isolate the peak containing z_{PEAK}, z_{WEIGHT} is defined as the weighted mean redshift within this peak. We restrict to a single peak in order to avoid confusion from bimodal and multimodal $p(z)$ such as those shown in bottom panels of Figure 1. For example, for a bimodal probability distribution a weighted mean calculated over both peaks would fall between the peaks, at a redshift where the probability is minimal. Restricting the weighting to a single peak ensures that the point estimate will fall in the region of maximum redshift probability.

A1 Point Estimate Metrics

We calculate the commonly used point estimate metrics of the overall photo-z scatter (σ_z, the standard deviation of the photo-z residuals), bias, and "catastrophic outlier rate". Specifically, we calculate the metrics as follows: we define e_z as

$$e_z = \frac{z_P - z_S}{1 + z_S} \tag{A1}$$

where z_P is the point estimate and z_S is the true redshift. In practice, because the standard deviation calculation is quite sensitive to the outliers, we define the photo-z scatter, σ in terms of the Interquartile Range (IQR), the difference between the 75th and 25th percentiles of the e_z distribution. In order to match the usual meaning of a 1σ interval, we scale the IQR and define $\sigma_{IQR} = IQR/1.349$, as there is a factor of 1.349 difference between the IQR and the standard deviation of a Normal distribution. While many other studies define the bias based on the *mean* offset between true and estimated redshift, in this study we define the bias as the median value of e_z for the sample. We use median as it is, once again, less sensitive to outliers than the mean. The catastrophic outlier fraction is defined as the fraction of galaxies with e_z greater than the *larger* of $3\sigma_{IQR}$ or 0.06, i.e. 3σ outliers with a floor of σ_{IQR}=0.02. For reference, the goals stated in Section 3.8 of the LSST Science Book (Abell et al. 2009) for photo-z performance in these metrics, assuming perfect training knowledge (as we are testing in this paper) are:

- RMS scatter$< 0.02(1+z)$
- bias < 0.003
- catastrophic outlier rate $< 10\%$

These definitions are similar, but not exactly the same, as the σ_{IQR} and median bias calculated here, but are similar enough for qualitative comparisons to the LSST goals.

Fig. A1 shows the point estimates for both z_{PEAK} and z_{WEIGHT}. Point density is shown with mixed contours to emphasize that most of the galaxies do fall close to the $z_{phot} = z_{spec}$ line, while blue points show differing characteristics of the outlier populations. The red dashed lines indicated the cutoff for catastrophic outliers, defined as: $max(0.06, 3\sigma_{IQR})$. As with the full $p(z)$ results, a variety of behaviours are evident in the different codes. Table A1 lists the scatter, bias, and catastrophic outlier fractions for the codes. The performance of the codes for point metrics is highly correlated with performance on $p(z)$ based tests, which is to be expected, given that the point-estimates were derived from the $p(z)$. Some discretization is evident in z_{PEAK}, particularly for SKYNET, due to the finite grid spacing of the reported $p(z)$. These discreteness effects are mitigated by the weighting of z_{WEIGHT}, resulting in a smoother distribution of redshift estimates. Several features perpendicular to the main $z_{phot} = z_{spec}$ line are evident. These features are due to the 4000 angstrom break passing through the gaps between adjacent LSST filters. These features are most prominent in template-based codes, but appear to some degree in all codes tested.

In even the best performing codes, there are visible occupied regions away from the $z_{phot} = z_{spec}$ line, corresponding to degenerate redshift solutions for certain LSST magnitudes and colors. While use of the full information available via $p(z)$ mitigates their impact, a full understanding of the outlier population is critical for LSST science, particularly in tomographic applications

Finally, we note that all twelve codes perform at or near the goals for point-estimates as outlined in the LSST Science Requirements Document[19] and Graham et al. (2018). This is to be expected, given that the requirements were designed such that a point estimate photo-z would meet these requirements for perfect training data to a depth of $i < 25$. But, it is still an encouraging sign, given an updated mock galaxy simulation and the expanded set of photo-z codes tested.

REFERENCES

Abbott T., et al., 2005, preprint (arXiv:astro-ph/0510346)
Abell P. A., et al., 2009, preprint (arXiv:0912.0201),
Aihara H., et al., 2018a, PASJ, 70, S4
Aihara H., et al., 2018b, PASJ, 70, S8
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
Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109
Benítez N., 2000, ApJ, 536, 571
Bernstein G., Huterer D., 2010, MNRAS, 401, 1399
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Blanton M. R., et al., 2005, AJ, 129, 2562
Bonnett C., 2015, MNRAS, 449, 1043
Bonnett C., 2016, Python wrapper to SkyNet, `https://pyskynet.readthedocs.io/en/latest/`
Bonnett C., et al., 2016, Phys. Rev. D, 94, 042005
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., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees, Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A
Brescia M., Cavuoti S., Amaro V., Riccio G., Angora G., Vellucci C., Longo G., 2018, preprint, (`arXiv:1802.07683`)
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Carrasco Kind M., Brunner R. J., 2014, MNRAS, 442, 3380
Cavuoti S., Amaro V., Brescia M., Vellucci C., Tortora C., Longo G., 2017, MNRAS, 465, 1959
Chen T., Guestrin C., 2016, in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. ACM, New York, NY, USA, pp 785–794, doi:10.1145/2939672.2939785, `http://doi.acm.org/10.1145/2939672.2939785`
Connolly A. J., et al., 2014, in Angeli G. Z., Dierickx P., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 9150, Modeling, Systems Engineering, and Project Management for Astronomy VI. p. 14, doi:10.1117/12.2054953
Dahlen T., et al., 2013, ApJ, 775, 93
Dawson W. A., Schneider M. D., Tyson J. A., Jee M. J., 2016, ApJ, 816, 11
Fernández-Soto A., Lanzetta K. M., Yahil A., 1999, ApJ, 513, 34
Firth A. E., Lahav O., Somerville R. S., 2003, MNRAS, 339, 1195
Freeman P. E., Izbicki R., Lee A. B., 2017, MNRAS, 468, 4556
Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, MNRAS, 441, 1741
Graham M. L., Connolly A. J., Ivezić Ž., Schmidt S. J., Jones R. L., Jurić M., Daniel S. F., Yoachim P., 2018, AJ, 155, 1
Green J., et al., 2012, preprint (arXiv:1208.4012),
Hildebrandt H., et al., 2010, A&A, 523, A31
Hunter J. D., 2007, Matplotlib: A 2D Graphics Environment, doi:10.1109/MCSE.2007.55
Ilbert O., et al., 2006, A&A, 457, 841
Ivezić Ž., et al., 2008, preprint (arXiv:0805.2366),
Izbicki R., Lee A. B., 2017, Electron. J. Statist., 11, 2800
Izbicki R., Lee A. B., Freeman P. E., 2017, Ann. Appl. Stat., 11, 698
Laureijs R., et al., 2011, preprint (1110.3193),
Leistedt B., Hogg D. W., 2017, ApJ, 838, 5
Malz A., Hogg D., in prep., CHIPPR, chippr
Malz A., Marshall P., DeRose J., Graham M., Schmidt S., Wechsler R., 2018, AJ, Accepted,
Mandelbaum R., et al., 2008, MNRAS, 386, 781
Massarotti M., Iovino A., Buzzoni A., 2001, A&A, 368, 74
Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes J. D., Castander F. J., Paltani S., 2017, ApJ, 841, 111
Newman J. A., 2008, ApJ, 684, 88
Newman J. A., et al., 2015, Astroparticle Physics, 63, 81
Oliphant T., 2007, Python for Scientific Computing, doi:10.1109/MCSE.2007.58
Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008, ApJ, 689, 709
Polsterer K. L., D'Isanto A., Gieseke F., 2016, preprint (arXiv:1608.08016),
Rasmussen C., Williams C., 2006, Gaussian Processes for Machine Learning. Adaptative computation and machine learning series, MIT Press, Cambridge, MA
Rau M. M., Seitz S., Brimioulle F., Frank E., Friedrich O., Gruen D., Hoyle B., 2015, MNRAS, 452, 3710
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, ApJ, 771, 30
Sadeh I., Abdalla F. B., Lahav O., 2016, PASP, 128, 104502
Sánchez C., Bernstein G. M., 2018, preprint, (`arXiv:1807.11873`)
Sánchez C., et al., 2014, MNRAS, 445, 1482
Schmidt M., 2005, minFunc: Unconstrained Differentiable Multivariate Optimization in Matlab, `http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`
Scott D. W., 1992, Multivariate Density Estimation. Theory, Practice, and Visualization. Wiley
Skrutskie M. F., et al., 2006, AJ, 131, 1163
Tanaka M., et al., 2018, PASJ, 70, S9
The LSST Dark Energy Science Collaboration et al., 2018, preprint, (`arXiv:1809.01669`)
Waskom M., et al., 2017, doi:10.5281/zenodo.824567
York D. G., et al., 2000, AJ, 120, 1579
de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, Exp. Astron., 35, 25
de Jong J. T. A., et al., 2017, A&A, 604, A134

[19] available at: `http://ls.st/srd`

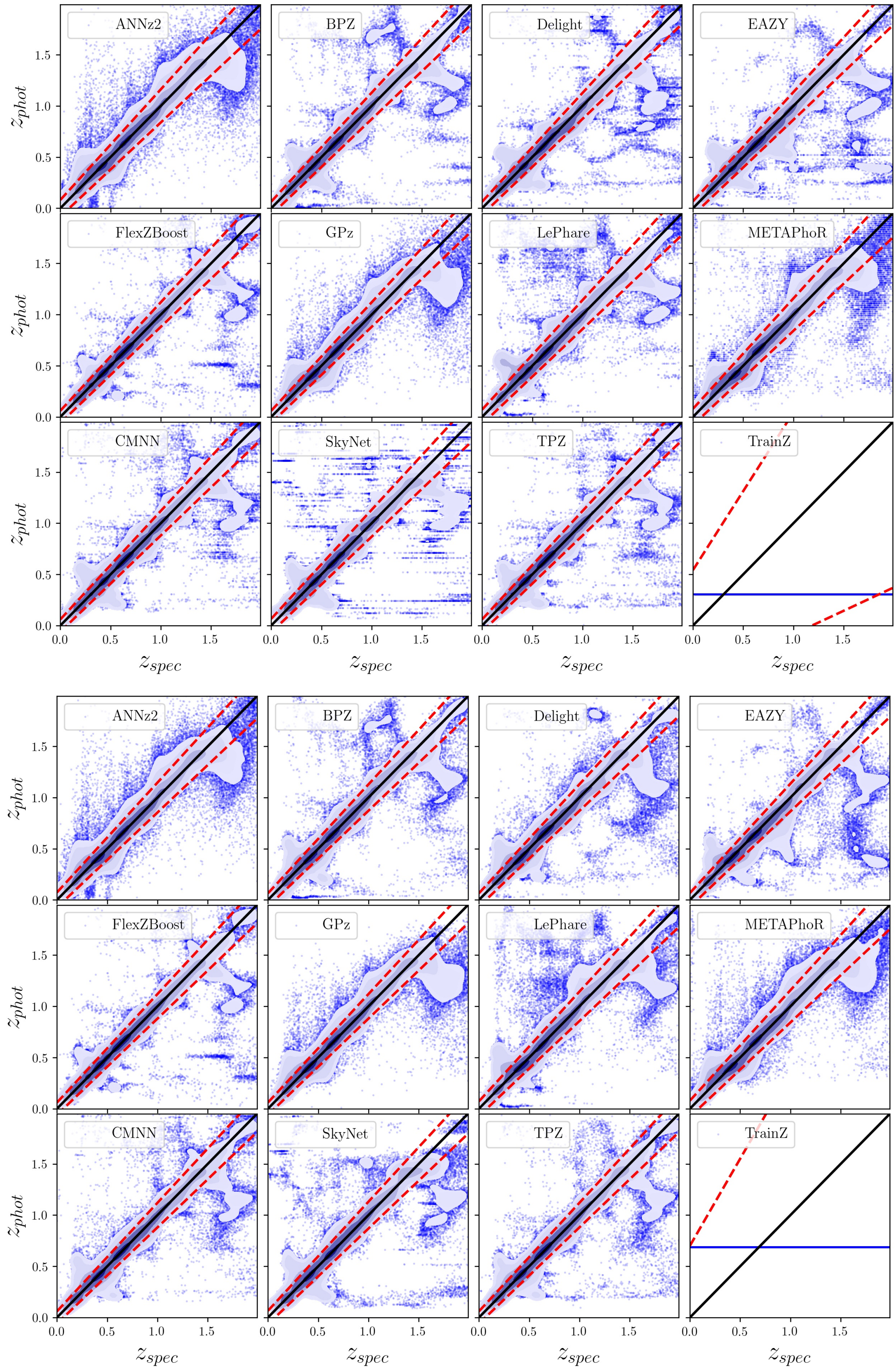

Figure A1. Point estimate photo-z's derived from the posteriors. Top panel shows z_{PEAK}, while bottom panel shows z_{WEIGHT}. Point estimate density is represented with fixed density contours, while outliers at lower density are represented by blue points. While use of point-estimate photo-z's is not recommended, they do make for useful comparative and visual diagnostics. In the lower-right panel of each plot, the TRAINZ estimator results in identical photo-z estimates at the mode and mean of the training set $N'(z)$ distribution for all galaxies.

Table A1. Point estimate statistics

	z_{PEAK}			z_{WEIGHT}		
Photo-z Code	$\frac{\sigma_{IQR}}{(1+z)}$	median	outlier fraction	$\frac{\sigma_{IQR}}{(1+z)}$	median	outlier fraction
ANNz2	0.0270	0.00063	0.044	0.0244	0.000307	0.047
BPZ	0.0215	-0.00175	0.035	0.0215	-0.002005	0.032
Delight	0.0212	-0.00185	0.038	0.0216	-0.002158	0.038
EAZY	0.0225	-0.00218	0.034	0.0226	-0.003765	0.029
FlexZBoost	0.0154	-0.00027	0.020	0.0148	-0.000211	0.017
GPz	0.0197	-0.00000	0.052	0.0195	0.000113	0.051
LePhare	0.0236	-0.00161	0.058	0.0239	-0.002007	0.056
METAPhoR	0.0264	0.00000	0.037	0.0262	0.001333	0.048
CMNN	0.0184	-0.00132	0.035	0.0170	-0.001049	0.034
Skynet	0.0219	-0.00167	0.036	0.0218	0.000174	0.037
TPZ	0.0161	0.00309	0.033	0.0166	0.003048	0.031
trainZ	0.1808	-0.2086	0.000	0.2335	0.022135	0.000