

# Implicit assumptions and their impact on photometric redshift PDF performance in the context of LSST

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

(LSST Dark Energy Science Collaboration)

<sup>1</sup> Department of Physics, University of California, One Shields Ave., Davis, CA, 95616, USA

<sup>2</sup> Center for Cosmology and Particle Physics, New York University, 726 Broadway, New York, 10003, USA

<sup>3</sup> Department of Physics, New York University, 726 Broadway, New York, 10003, USA

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

<sup>5</sup> INAF-Capodimonte Observatory, Salita Moiariello 16, I-80131, Napoli, Italy

<sup>6</sup> Department of Physics E. Pancini, University Federico II, via Cinthia 6, I-80126, Napoli, Italy

<sup>7</sup> King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

<sup>8</sup> Information Engineering, Parks Road, Oxford, OX1 3PJ, UK

<sup>9</sup> Department of Astronomy, University of Washington, Box 351580, U.W., Seattle WA 98195, USA

<sup>10</sup> Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS, Montpellier, France

<sup>11</sup> Department of Statistics & Data Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

<sup>12</sup> Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854-8019 USA

<sup>13</sup> Department of Physics, University of Washington, Box 351560, Seattle, WA 98195, USA

<sup>14</sup> Argonne National Laboratory, Lemont, IL 60439, USA

<sup>15</sup> Department of Physics and Astronomy and PITT PACC, University of Pittsburgh, 3941 O'Hara St., Pittsburgh, PA 15260, USA

<sup>16</sup> Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK

<sup>17</sup> Department of Physics and Astronomy, University of the Western Cape, Bellville 7535, South Africa

<sup>18</sup> Department of Statistics, Federal University of Sao Carlos, Sao Carlos, Brazil

<sup>19</sup> External collaborator

27 November 2018

## ABSTRACT

Photometric redshift (photo- $z$ ) probability distribution functions (PDFs) are a key data product of nearly all upcoming galaxy imaging surveys. However, the photo- $z$  PDFs resulting from different techniques are not in general consistent with one another, and an optimal method for obtaining an accurate PDF remains unclear. We present the results of an initial study of the Large Synoptic Survey Telescope Dark Energy Science Collaboration (LSST-DESC), the first in a planned series of papers testing multiple photo- $z$  codes on simulations of upcoming LSST galaxy photometry catalogues. This initial test evaluates photo- $z$  algorithms in the presence of representative training data and in the absence of several common sources of systematic errors that affect the procedures by which photo- $z$  PDFs are derived. The photo- $z$  PDFs are evaluated using multiple metrics and we observe several trends, including an overall over/under-prediction in the broadness of the PDFs for several of the codes. A careful accounting of all systematics discovered 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

## 2 LSST Dark Energy Science Collaboration

### 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 galaxies, expected to number several billion 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)<sup>1</sup> lists the individual galaxy photometric redshift goals for a magnitude limited sample with  $i < 25$  as: root-mean-square error with a goal of  $\sigma_z < 0.02(1+z)$ ;  $3\sigma$  “catastrophic outlier” rate below 10%; bias below 0.003<sup>2</sup>. 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 initial 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  $\chi^2$  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.

<sup>1</sup> available at <https://docushare.lsstcorp.org/docushare/dsweb/Get/LPM-17>

<sup>2</sup> 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.

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-

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 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)<sup>3</sup> 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 a 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 put in cites to in prep Buzzard papers catalogue construction started with a dark matter only simulation. This N-body simulation contained  $2048^3$  particles in a  $400 \text{ Mpc h}^{-1}$  box. [N] 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.)<sup>4</sup> procedure was used. This consisted of training an empirical relation between absolute  $r$ -band magnitude, local galaxy density, and SED using a sample of  $\sim 5e^5$  galaxies from the magnitude-limited Sloan Digital Sky Survey Data Release 6 Value Added Galaxy Catalog (Blanton et al. 2005)[Note: is this the proper reference to SDSS-NYU VAGC? File is called combined\_dr6\_cooper.fits, but I don't see which Cooper et al 2006 this is supposed to refer to?]. Each SDSS spectrum is fit with a sum of five SED components using the K-CORRECT v? software package<sup>5</sup> (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 “random” [details] galaxy with “similar” [details] 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. 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.

#### 2.1.1 Selection of training and test sets

The total catalogue covered 400 square degrees and contained 238 million galaxies to an apparent magnitude limit

<sup>3</sup> Available at: [http://lsst-desc.org/sites/default/files/DESC\\_SRM\\_V1\\_1.pdf](http://lsst-desc.org/sites/default/files/DESC_SRM_V1_1.pdf)

<sup>4</sup> <https://github.com/vipasu/addseds>

<sup>5</sup> <http://kcorrect.org>

## 4 LSST Dark Energy Science Collaboration

of  $r = 29$  and spanning the redshift range  $0 < z \leq 8.7$ . This catalogue contained two orders of magnitude more galaxies than were needed for this study, so only  $\sim 8$  square degrees were used. Systematic problems with galaxy colors above  $z > 2$  were observed, so the catalogue was trimmed 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. The resulting catalogues contain 111 171 training galaxies and 1 000 883 test galaxies. We restrict our analysis to a sample with  $i < 25.3$ , 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.

### 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  $\sim 5e^5$  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  $\sim 500\,000$  galaxies in the SDSS sample and run a simple K-means clustering algorithm on this five dimensional space. The K-means cluster centres span the space of coefficients and properly reflect the underlying density in the coefficient space, thus providing a reasonable approximation for a 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, however not every template code uses this set of one hundred templates: 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.

### 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 simulations 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. 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.

## 3 METHODS

Here we outline the photo- $z$  PDF codes tested in this study. In total, eleven distinct codes are tested. This sample is not comprehensive, but does cover a broad range of current-generation codes. 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.

The questions that must be answered for each code are: 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

#### 3.1.1 BPZ

BPZ<sup>6</sup> (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 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):

$$p(z|C, m_0) \propto \sum_T p(z, T|m_0) p(C|z, T) \quad (1)$$

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

<sup>6</sup> <http://www.stsci.edu/~dcoe/BPZ/>

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

Code	Type	Paper	Website
BPZ	template	Benítez (2000)	<a href="http://www.stsci.edu/~dcoe/BPZ/">http://www.stsci.edu/~dcoe/BPZ/</a>
EAZY	template	Brammer et al. (2008)	<a href="https://github.com/gbrammer/eazy-photoz">https://github.com/gbrammer/eazy-photoz</a>
LEPHARE	template	Arnouts et al. (1999)	<a href="http://www.cfht.hawaii.edu/~arnouts/lephare.html">http://www.cfht.hawaii.edu/~arnouts/lephare.html</a>
ANNz2	ML	Sadeh et al. (2016)	<a href="https://github.com/IftachSadeh/ANNz2">https://github.com/IftachSadeh/ANNz2</a>
DELIGHT	ML/template	Leistedt & Hogg (2017)	<a href="https://github.com/ixxkael/Delight">https://github.com/ixxkael/Delight</a>
FLEXZBOOST	ML	Izbicki & Lee (2017)	<a href="https://github.com/tospisici/flexcode">https://github.com/tospisici/flexcode;</a> <a href="https://github.com/rizbicki/FlexCoDE">https://github.com/rizbicki/FlexCoDE</a>
GPz	ML	Almosallam et al. (2016b)	<a href="https://github.com/OxfordML/GPz">https://github.com/OxfordML/GPz</a>
METAPHOR	ML	Cavuoti et al. (2017a)	<a href="http://dame.dsfa.unina.it">http://dame.dsfa.unina.it</a>
CMNN	ML	Graham et al. (2018)	-
SKYNET	ML	Graff et al. (2014)	<a href="http://ccpforge.cse.rl.ac.uk/gf/project/skynet/">http://ccpforge.cse.rl.ac.uk/gf/project/skynet/</a>
TPZ	ML	Carrasco Kind & Brunner (2013)	<a href="https://github.com/mgckind/MLZ">https://github.com/mgckind/MLZ</a>
TRAINZ	N/A	See Section 3.3	

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 with negative flux in a measured band, the placeholder value is replaced with an estimate one  $\sigma$  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<sup>7</sup> (Easy and Accurate Photometric Redshifts from Yale, Brammer et al. 2008) is a template-based photo-z code that includes several features that improve on the basic  $\chi^2$  fit used in many template codes. It 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  $\chi^2$ . The minimized  $\chi^2(z)$  is then combined with an apparent magnitude prior to obtain the posterior redshift probability distribution, although some argue that this is not the mathematically correct way of calculating the posteriors. 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 are calculated on a 200-point redshift grid spanning  $0 \leq z \leq 2$ , and include the application of a type-independent apparent magnitude prior estimated from the training data.

### 3.1.3 LePhare

LEPHARE<sup>8</sup> (Photometric Analysis for Redshift Estimate, Arnouts et al. 1999; Ilbert et al. 2006) is a photo-z reconstruction code based on a  $\chi^2$  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. LEPHARE has been used to produce the COSMOS2015 photo-z catalogue (Laigle et al. 2016).

Each SED is redshifted in steps of  $\Delta z = 0.01$  and convolved with the simulated LSST filter transmission curves (accounting for instrument efficiency). The opacity of the inter-galactic medium has been set to zero as no additional reddening has been included in the Buzzard simulations. The computed photo-z is then the value that minimizes the merit function  $\chi^2(z, T, A)$  from Arnouts et al. (1999):

$$\chi^2(z, T, A) = \sum_f^{N_f} \left( \frac{F_{\text{obs}}^f A \times F_{\text{pred}}^f(T, z)}{\sigma_{\text{obs}}^f} \right)^2 \quad (2)$$

where  $A$  is a normalization factor,  $F_{\text{pred}}^f(T, z)$  is the flux predicted for a template  $T$  at redshift  $z$ .  $F_{\text{obs}}^f$  is the observed flux in a given band  $f$  and  $\sigma_{\text{obs}}^f$  the associated observational error. The index  $f$  refers to the considered band and  $N_f$  is the total number of filters.

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. The full  $p(z)$  corresponds to the likelihoods calculated at each point on our  $z$ -grid.

## 3.2 Training-based Codes

### 3.2.1 ANNz2

ANNz2<sup>9</sup> (Sadeh et al. 2016) is a powerful package that has the ability to employ several machine learning algorithms, including artificial neural networks (ANN), boosted

<sup>7</sup> <https://github.com/gbrammer/eazy-photoz>

<sup>8</sup> <http://www.cfht.hawaii.edu/~arnouts/lephare.html>

<sup>9</sup> <https://github.com/IftachSadeh/ANNz2>

## 6 LSST Dark Energy Science Collaboration

402 decision tree (BDT) and k-nearest neighbour (KNN). Us-  
 403 ing the Toolkit for Multivariate Data Analysis (TMVA)  
 404 with ROOT<sup>10</sup>, it can run multiple machine learning algo-  
 405 rithms for a single training and outputs photo- $z$ 's based on  
 406 a weighted average of their performances.

407 ANNz2 is capable of producing both photo- $z$  point es-  
 408 timates and redshift posterior probability distributions  $p(z)$ ,  
 409 it could also conduct classifications and supports reweight-  
 410 ing between samples. The PDFs are produced by propa-  
 411 gating the intrinsic uncertainty on the input parameters  
 412 and the uncertainty in the machine learning method to  
 413 the expected photo- $z$  solution, averaged over multiple runs  
 414 weighted based on the performance of each run. ANNz2  
 415 presents its photo- $z$  uncertainty different from many codes  
 416 by using the KNN method: it estimates the photo- $z$  bias be-  
 417 tween each object and a fixed number of nearest neighbours  
 418 in parameter space, it then takes the 68th percentile width of  
 419 the distribution of the bias. This is based on the implication  
 420 that objects with similar photometric properties would have  
 421 similar uncertainties, and therefore the photometric errors  
 422 of the inputs are not propagated into the code.

423 In this study, ANNz2 v. 2.0.4 was used. The full PDF  
 424 for each galaxy is also produced with a linear stepsize of  
 425  $z = 0.01$  for  $0 \leq z \leq 2$ . A set of 5 ANNs with architec-  
 426 ture  $6 : 12 : 12 : 1$  (6 *ugrizy* inputs, 2 hidden layers with  
 427 12 nodes each, and 1 output) with different random seeds  
 428 are used during each training. Half of the training set is  
 429 used as a validation set to prevent overtraining. All training  
 430 objects are set to have detected magnitudes, however the  
 431 non-detections ( $\text{mag} = -99$ ) in the testing set are replaced  
 432 with the mean of that particular band.

458 photometric redshift estimator described in G18 to the sim-  
 459 ulated data. Compared to its application in G18, there are  
 460 some minor differences in the application of this estimator  
 461 to the Buzzard catalogue. First, we do not impose non-  
 462 detections on galaxies with a magnitude fainter than the ex-  
 463 pected LSST 10-year limiting magnitude or bright enough to  
 464 saturate with LSST: *all* of the photometry for all the galax-  
 465 ies in the test and training sets are used for this experiment.  
 466 Second, as in G18 we do apply an initial cut in color to  
 467 the training set before calculating the Mahalanobis distance  
 468 in order to accelerate processing, and also use a magnitude  
 469 pseudo-prior to improve photo- $z$  estimates, but for both we  
 470 have used different cut-off values that are appropriate for the  
 471 Buzzard galaxies' colors and magnitudes. Third, we set dif-  
 472 ferent parameters for the identification of the color-matched  
 473 subset of training galaxies and the selection of a photometric  
 474 redshift estimate. In G18 we used a percent point function  
 475 (PPF) value of 0.68 to identify the color-matched subset of  
 476 training galaxies and used the redshift of nearest neighbour  
 477 in color-space as the photo- $z$  estimate. These choices work  
 478 well when the desire is to obtain accurate photo- $z$  estimates  
 479 for most test-set galaxies, but does not return a robust  $p(z)$   
 480 in all cases – especially for galaxies that are bright and/or  
 481 have few matches in color-space. Since a robust estimate  
 482 of  $p(z)$  is desired for this work we make several changes to  
 483 our implementation of the CMNN photo- $z$  estimator. We  
 484 continue to use a percent point function of PPF = 0.95 to  
 485 generate the subset of color-matched training galaxies, but  
 486 weight them by the inverse of their Mahalanobis distance.  
 487 This weighting maintains some of the accuracy that was pre-  
 488 viously achieved by simply using the nearest neighbour in  
 489 color-space. We then use the weights to create the  $p(z)$  in-  
 490 stead of having the redshift of each color-matched training-  
 491 set galaxy count equally. To obtain a robust estimate of the  
 492  $p(z)$  for galaxies with a small number of color-matched train-  
 493 ing set galaxies, when this number is less than 20 the nearest  
 494 20 neighbours in color-space are used instead, and we con-  
 495 volve the  $p(z)$  with a Gaussian with a standard deviation  
 496 of:

$$497 \sigma = \sigma_{\text{train}} \sqrt{(\text{PPF}_{20}/0.95)^2 - 1} \quad (3)$$

498 to appropriately broaden it so that the  $p(z)$  for these test  
 499 galaxies represents the enlarged PPF value associated with  
 500 it. Overall, these three changes will yield poorer accuracy  
 501 photo- $z$  compared to those presented in G18, but they will  
 502 all have significantly more robust estimates of the  $p(z)$ , par-  
 503 ticularly for the brightest test galaxies. This is sufficient  
 504 for this work because, as described in G18, the goal of the  
 505 CMNN photo- $z$  estimator was never to provide the “best”  
 506 (or even competitive) estimates in the first place, given its  
 507 reliance on a deep training set, but rather to provide a means  
 508 for direct comparisons between LSST photometric quality  
 509 and photo- $z$  estimates. With this work we show how the in-  
 510 put parameters should be set in order to return robust  $p(z)$   
 511 estimates in addition to point value estimates.

457 We have applied the nearest-neighbours color-matching

<sup>10</sup> <http://tmva.sourceforge.net/>

512 

### 3.2.3 Delight

513 DELIGHT<sup>11</sup> (Leistedt & Hogg 2017) infers photo-z’s by using  
 514 a data-driven model of latent SEDs and a physical model  
 515 of photometric fluxes as a function of redshift. Generally,  
 516 machine learning methods rely on representative training  
 517 data with similar band passes, while template based meth-  
 518 ods rely on a complete library of templates based on phys-  
 519 ical models constructed. DELIGHT is constructed in attempt  
 520 to combine the advantages and eliminate the disadvantages  
 521 of both template-based and machine learning algorithms: it  
 522 constructs a large collection of latent SED templates (or  
 523 physical flux-redshift models) from training data, with a  
 524 template SED library as a guide to the learning of the model.  
 525 The advantage of DELIGHT is that it neither needs represen-  
 526 tative training data in the same photometric bands, nor does  
 527 it need detailed galaxy SED models to work.

528 This conceptually novel approach is done by using  
 529 Gaussian processes operating in flux-redshift space. The pos-  
 530 terior distribution on the redshift of a target galaxy is ob-  
 531 tained via a pairwise comparison with training galaxies,

$$532 p(z|\hat{\mathbf{F}}) \approx \sum_i p(\hat{\mathbf{F}}|z, t_i) p(z|t_i) p(t_i), \quad (4)$$

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

$$538 p(\hat{\mathbf{F}}|z, t_i) = \int p(\hat{\mathbf{F}}|\mathbf{F}) p(\mathbf{F}|z, z_i, \hat{\mathbf{F}}_i) d\mathbf{F}, \quad (5)$$

539 where  $p(\hat{\mathbf{F}}|\mathbf{F})$  is the likelihood function, it compares the  
 540 noisy real flux  $\hat{\mathbf{F}}$  with the noiseless flux  $\mathbf{F}$  obtained from the  
 541 linear combination of template models, carefully constructed  
 542 to account for model uncertainties and different normaliza-  
 543 tion of the same SED; while  $p(\mathbf{F}|z, z_i, \hat{\mathbf{F}}_i)$  is the prediction  
 544 of flux at a different redshift  $z$  with respect to the training  
 545 object with redshift  $z_i$  and flux  $\hat{\mathbf{F}}_i$ . Eq. 5 is essentially the  
 546 probability that the training and the target galaxies having  
 547 the same SED but at a different redshift. The flux prediction  
 548  $p(\mathbf{F}|z, z_i, \hat{\mathbf{F}}_i)$  of the training galaxy at redshift  $z$  is modeled  
 549 via a Gaussian process,

$$550 F_b \sim \mathcal{GP}(\mu^F, k^F), \quad (6)$$

551 with mean function  $\mu^F$  and kernel  $k^F$ , both imposed to  
 552 capture expected correlations resulting from the known un-  
 553 derlying physics (i.e., fluxes resulting from observing SEDs  
 554 through filter response, and the SEDs being redshifted). The  
 555 reader should refer to Leistedt & Hogg (2017) for further de-  
 556 tails.

557 In this study, all 100 ordered Buzzard templates, as  
 558 described in Section 2.1.2, were used in DELIGHT, and the  
 559 Gaussian process was trained with a subset of 50 000 galaxies.  
 560 Photometric uncertainties from the inputs are propa-  
 561 gated into the code, while non-detections for each band are  
 562 set to the mean of the respective bands. Default settings

563 of DELIGHT were use, with the exception that the PDF bins  
 564 were set to be linear instead of logarithmic, with 200 equally-  
 565 spaced bins between  $0.0 < z < 2.0$ . In this study a flat prior  
 566 is assumed.

567 

### 3.2.4 FlexZBoost

568 FLEXZBOOST<sup>12</sup> (Izbicki & Lee 2017) is a particular realiza-  
 569 tion of **FlexCode**, which is a general-purpose methodology  
 570 for converting any conditional mean point estimator of  $z$  to  
 571 a conditional density estimator  $f(z|\mathbf{x})$ , where  $\mathbf{x}$  here repre-  
 572 sents our photometric covariates and errors.<sup>13</sup> The key idea  
 573 is to expand the unknown function  $f(z|\mathbf{x})$  in an orthonormal  
 574 basis  $\{\phi_i(z)\}_i$ :

$$575 f(z|\mathbf{x}) = \sum_i \beta_i(\mathbf{x}) \phi_i(z). \quad (7)$$

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

$$578 \beta_i(\mathbf{x}) = \mathbb{E}[\phi_i(z)|\mathbf{x}] \equiv \int f(z|\mathbf{x}) \phi_i(z) dz. \quad (8)$$

579 These coefficients can easily be estimated from data by re-  
 580 gression.

581 In this paper, we use XGBOOST (Chen & Guestrin 2016)  
 582 for the regression part as these techniques scale well for mas-  
 583 sive data; it should however be noted that FLEXCODE-RF  
 584 (also on GitHub), based on Random Forests, generally per-  
 585 forms better for smaller data sets. As our basis, we choose a  
 586 standard Fourier basis. There are two tuning parameters in  
 587 our  $p(z)$  estimate: (i) the number of terms,  $I$ , in the series  
 588 expansion in Eq. 7, and (ii) an exponent  $\alpha$  that we use to  
 589 sharpen the computed density estimates  $\hat{f}(z|\mathbf{x})$ , according  
 590 to  $\tilde{f}(z|\mathbf{x}) \propto \hat{f}(z|\mathbf{x})^\alpha$ . We split the “train data” into a train-  
 591 ing set (85%) and a validation set (15%), and choose both  $I$   
 592 and  $\alpha$  in an automated way by minimizing the weighted  $L_2$ -  
 593 loss function (Eq. 5 in Izbicki & Lee 2017) on the validation  
 594 set.

595 Although **FlexCode** offers a *lossless compression* of the  
 596 photo-z estimates (in this study, one can reconstruct  $\tilde{f}(z|\mathbf{x})$   
 597 exactly at any resolution from estimates of the first 35 co-  
 598 efficients, Eq. 8, for a Fourier basis  $\{\phi_i(z)\}_i$ ), we discretize  
 599 our final estimates into 200 bins linearly spaced in  $0 < z < 2$   
 600 for easy comparison with other algorithms. Using a higher  
 601 resolution may yield better results (with no added cost in  
 602 storage).

603 

### 3.2.5 GPz

604 GPz<sup>14</sup> (Almosallam et al. 2016a,b) is a sparse Gaussian pro-  
 605 cess based code, a fast and a scalable approximation of full  
 606 Gaussian Processes (Rasmussen & Williams 2006), with the  
 607 added feature of being able to produce input-dependent vari-  
 608 ance estimations (heteroscedastic noise). The model assumes

<sup>12</sup> <https://github.com/tpospisi/flexcode>;  
<https://github.com/rizbicki/FlexCoDE>

<sup>13</sup> Instead of  $p(z)$ , we use the notation  $f(z|\mathbf{x})$  to explicitly show the dependence on  $\mathbf{x}$ .

<sup>14</sup> <https://github.com/OxfordML/GPz>

<sup>11</sup> <https://github.com/ixkael/Delight>

## 8 LSST Dark Energy Science Collaboration

that the probability of the output  $y$ , the redshift, given the input  $x$ , the photometry, is  $p(y|x) = \mathcal{N}(y|\mu(x), \sigma(x)^2)$ . The mean function,  $\mu(x)$ , and the variance function  $\sigma(x)^2$  are both linear combinations of basis functions that take the following form:

$$f(x) = \sum_{i=1}^m \phi_i(x) w_i, \quad (9)$$

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  $|z_{\text{spec}} - z_{\text{phot}}|/(1 + z_{\text{spec}})$  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 the missing values using a simple linear model that estimates the missing variables 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  $\mu$  and its covariance  $\Sigma$ , then we use the following equation to estimate the missing values from the observed ones

$$x_u = \mu_u + \Sigma_{uo} \Sigma_{oo}^{-1} (x_o - \mu_o), \quad (10)$$

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  $\mu$  and  $\Sigma$ ). 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. 2017a) is a pipeline designed to provide photo-z's 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), already validated on photo-z's in several cases (de Jong et al. 2017; Cavuoti et al. 2017b, 2015; Brescia et al. 2014, 2013; Biviano et al. 2013). Due to its plug-in based modular nature, METAPHOR can be easily replaced by any other photo-z prediction kernel, regardless its implementation, by taking the I/O interface compliance as unique constrain.

At a higher level, the pipeline mainly consists of three modules: (i) *data pre-processing*, including a catalogue cross-matching sub-module (based on the tool C3, Riccio et al. 2017), a sub-module for photometric evaluation and error estimation of the multi-band catalogue used as Knowledge Base (KB), and a sub-module dedicated to the perturbation of the photometric KB, propaedeutic to the PDF estimation; (ii) *photo-z prediction*, which is the training/validation/test phase, producing the photo-z's point estimates, based on a pre-selected ML method; (iii) *PDF estimation*, specifically designed to calculate the PDF of the photo-z estimation errors. The last module includes also a post-processing tool, providing some statistics on the produced point estimates and PDFs.

The photometry perturbation law is based on the formula  $m_{ij} = m_{ij} + \alpha_i F_{ij} * u_{\mu=0, \sigma=1}$ , where  $\alpha_i$  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 introduced, the photo-z point estimate prediction engine of METAPHOR is based on the MLPQNA model, whose photo-z regression training error, used by the quasi Newton learning rule, is based on the least square error and Tikhonov  $L_2$ -norm regularization (Hofmann & Mathé 2018).

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 fill the missing detection, it goes without saying that for such points the reliability of PDFs and point estimation is lower. No cross validation has been used.

726 **3.2.7 SkyNet**

727 SKYNET<sup>15</sup> (Graff et al. 2014) is a publicly available neural  
 728 network software, based on a 2nd order conjugate gradient  
 729 optimization scheme (see Graff et al. 2014, for further de-  
 730 tails). It has been used efficiently for redshift PDF estimates  
 731 (Sánchez et al. 2014; Bonnett 2015; Bonnett et al. 2016).

732 The neural network is configured as a standard multi-  
 733 layer perceptron with three hidden layers and one input layer  
 734 with 12 nodes (the 6 magnitudes and their errors). The clas-  
 735 sifier is laid out such that the hidden layers have 20:40:40  
 736 nodes each, all rectified linear units, and the output layer  
 737 has 200 nodes (corresponding to 200 bins for the PDF) acti-  
 738 vated with a “softmax” function so that they automatically  
 739 sum to 1.

740 To avoid over-fitting, a 30 per cent fraction of the train-  
 741 ing set is used as validation, and the training is stopped as  
 742 soon as the error rate begins to increase in the validation  
 743 set. The weights are randomly initialized based on normal  
 744 sampling. The error function is a standard chi-square func-  
 745 tion for the regressor, and a cross-entropy function for the  
 746 classifier. Finally, the data are all whitened before process-  
 747 ing, with magnitudes pegged to (45,45,40,35,42,42) and their  
 748 errors pegged to (20,20,10,5,15,15) for *ugrizy* filters, respec-  
 749 tively.

750 **3.2.8 TPZ**

751 TPZ<sup>16</sup> (Trees for Photo-*z*, Carrasco Kind & Brunner 2013;  
 752 Carrasco Kind & Brunner 2014) is a parallel machine learning  
 753 algorithm that generates photometric redshift PDFs us-  
 754 ing prediction trees and random forest techniques. The code  
 755 recursively splits the input data (i. e. the training sample),  
 756 into two branches, one after another, until a terminal leaf is  
 757 created that meets a termination criterion (e. g. a minimum  
 758 leaf size or a variance threshold). Bootstrap samples from  
 759 the training data and associated errors are used to build a  
 760 set of prediction trees. In order to minimize correlation be-  
 761 tween the trees, the data is divided in such a way that the  
 762 highest information gain among the random subsample of  
 763 features is obtained at every point. The regions in each ter-  
 764 minal leaf node corresponds to a specific subsample of the  
 765 entire data that possesses similar properties.

766 The training data is examined before running TPZ.  
 767 Since TPZ does not handle non-detections (magnitudes  
 768 flagged as 99.0), we replace these values with an approxi-  
 769 mation of the  $1\sigma$  detection threshold, i. e. a signal to noise  
 770 ratio of 1 in terms of magnitude uncertainty using the equa-  
 771 tion  $dm = 2.5 \log(1 + N/S)$  where  $dm \sim 0.7526$  mag  
 772 for  $N/S = 1$ . That is, for each band, we replace the non-  
 773 detection with the magnitude corresponding to the error of  
 774 0.7526 from the error model forecasted for 10-year LSST  
 775 data. The Out-of-Bag (Breiman et al. 1984; Carrasco Kind  
 776 & Brunner 2013) cross-validation technique is used within  
 777 TPZ to evaluate its predictive validity and determine the  
 778 relative importance of the different input attributes. We em-  
 779 ployed this information to calibrate our algorithm.

780 In the present work, the LSST magnitudes  $u$ ,  $g$ ,  $r$ ,  $i$   
 781 and colors  $u-g$ ,  $g-r$ ,  $r-i$ ,  $i-z$ ,  $z-y$  and their associated

782 errors are used in the process of growing 100 trees with a  
 783 minimum leaf size of 5 (the  $z$  and  $y$  magnitudes did not  
 784 show significant correlation with the redshift in our cross-  
 785 validation, so we did not use them when constructing our  
 786 trees). We partitioned our redshift space into 100 bins from  
 787  $z = 0.005$  to  $z = 2.0$  and smoothed each individual PDF  
 788 with a smoothing scale of twice the bin size.

789 **3.3 Simple Ensemble Estimator**

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

$$793 p(z) = \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} z_{train} \quad (11)$$

794 That is, we simply set the redshift PDF of every galaxy equal  
 795 to the normalized  $N(z)$  of the training sample. As the train-  
 796 ing sample is drawn from the same underlying distribution  
 797 as the test sample, modulo small deviations due to sam-  
 798 ple size, the quantiles of the training and test distributions  
 799 should be identical. This is a wildly unrealistic estimator, as  
 800 it assigns all galaxies, no matter their apparent magnitude,  
 801 colour, or true redshift, the same redshift PDF, and is thus  
 802 uninformative at the level of individual object redshifts, but  
 803 is designed to perform very well for the ensemble of all ob-  
 804 jects. We will discuss this method and cautions relative to  
 805 metrics in Section 5.3.

806 **4 METRICS FOR QUANTIFYING PDF  
 807 COMPARISONS**

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

$$815 \hat{p}_j(z_i) = p(z|d_i, I_D, I_H), \quad (12)$$

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

827 In this study, we isolate the differences in prior infor-  
 828 mation specific to each method by using a single training set  
 829  $I_D^{ML}$  for all machine learning-based codes and a single tem-  
 830 plate library  $I_D^T$  for all template-based codes, and these sets

15 <http://ccpforge.cse.rl.ac.uk/gf/project/skynet/>

16 <https://github.com/mgckind/MLZ>

831 of prior information are carefully constructed to be represen-  
 832 tative and complete, we have  $I_D^{ML} \equiv I_D^T$  for every method  
 833  $H$ . Thus, we are saying

$$834 \quad \frac{\hat{p}_{i,H}(z)}{\hat{p}_{i,H'}(z)} \approx \frac{p(z|d_i, I_H)}{p(z|d_i, I_{H'})}, \quad (13)$$

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

839 As mentioned previously, there are cosmology probes  
 840 that require knowledge of individual galaxy  $p(z)$  and others  
 841 that require only knowledge of the ensemble redshift distri-  
 842 bution,  $N(z)$ . Due to the paucity of principled techniques  
 843 for using and validating photo- $z$  PDFs, there are few alter-  
 844 natives to the common practice of reducing photo- $z$  PDFs  
 845 to point estimates. Though this practice should not be en-  
 846 couraged, we also calculate traditional metrics based on the  
 847 most common point estimators derived from photo- $z$  PDFs.  
 848 Those seeking to establish a connection to traditional ways  
 849 of thinking about redshift estimation may consult the Ap-  
 850 pendix for these results.

851 There are a number of metrics that can be used to test  
 852 the accuracy of a photo- $z$  interim posterior as an estimator  
 853 of a true photo- $z$  posterior if it is known. Even for simulated  
 854 data, the true photo- $z$  PDF is in general not accessible un-  
 855 less the redshifts are in fact drawn from the true photo- $z$   
 856 PDFs, a mock catalogue generation procedure that has not  
 857 yet appeared in the literature. Furthermore, only limited ap-  
 858 plications of photo- $z$  PDFs that could be used as the basis  
 859 for a metric have been presented in the literature. The most  
 860 popular application by far is the calculation of the overall  
 861 redshift distribution  $N(z)$ , the true value of which is known  
 862 for the BUZZARD simulation and will be denoted as  $N'(z)$ .  
 863 Though alternatives exist (Malz & Hogg prep), stacking ac-  
 864 cording to

$$865 \quad \hat{N}^H(z) \approx \frac{1}{N_{tot}} \sum_i^{N_{tot}} \hat{p}_i^H(z) \quad (14)$$

866 is the most widely accepted method for estimating the red-  
 867 shift distribution from photo- $z$  PDFs. If we assume that  
 868 the response of estimators of  $N(z)$  is uniform across all ap-  
 869 proaches  $H$ , then we may interpret metrics on the accuracy  
 870 of  $\hat{N}(z)$  obtained in this way. We must note, however, that  
 871 this is a poor assumption in general. Under the setup of this  
 872 paper, the true redshift distribution  $N'(z) = p(z|I_D)$  (i.e.  
 873 because our training data is representative, the interim prior  
 874 is the truth). In this ideal case, the method that would give  
 875 the best approximation to  $N'(z)$  would be one that neglects  
 876 all the information contained in the photometry  $\{d_i\}_{N_{tot}}$   
 877 and gives every galaxy the same photo- $z$  PDF  $\hat{p}_i(z) = N'(z)$   
 878 for all  $i$ . This is the exact estimator, TRAINZ, that we have  
 879 described in Section 3.3, and which will serve as a point of  
 880 reference for the other codes.

881 The exact implementation of the stacked estimator  
 882  $\hat{N}^H(z)$  will depend on the parametrization of the photo- $z$   
 883 PDFs, which may differ across codes and can affect the pre-  
 884 cision of the estimator (Malz et al. 2018); even considering a  
 885 single method under the same parametrization, say a piece-  
 886 wise constant function over bins or a set of samples from

887 the posterior, an estimator using  $2N$  bins or samples will  
 888 trivially be more precise than an estimator using  $N$  bins or  
 889 samples. In order to minimize the effects of such choices,  
 890 we asked those running all eleven codes to output  $p(z)$  pa-  
 891 rameterized with a generous  $\approx 200$  piecewise constant bins  
 892 spanning  $0 < z < 2$ . The piecewise constant format is chosen  
 893 because of its established presence in the literature, and the  
 894 choice of 200 bins was motivated by the approximate number  
 895 of columns expected to be available for storage of  $p(z)$  for  
 896 the final LSST Project tables.<sup>17</sup> All  $p(z)$  catalogues are pro-  
 897 cessed using the QP software package (Malz et al. 2018)<sup>18</sup>  
 898 for manipulating and calculating metrics of 1-dimensional  
 899 PDFs. We will discuss the choice of  $p(z)$  parameterization  
 900 further in Section 5.

## 901 4.1 Metrics of an ensemble of photo- $z$ interim 902 posteriors

### 903 4.1.1 Probability integral transform (PIT)

904 The probability integral transform (PIT) (Polsterer et al.  
 905 2016) is defined for each individual galaxy as:

$$906 \quad \text{PIT} = \int_{-\infty}^{z_{\text{true}}} p(z) dz. \quad (15)$$

907 The distribution of PIT values quantifies the behavior of the  
 908 ensemble of photo- $z$  PDFs, enabling us to evaluate whether  
 909 the  $p(z)$  is, on average, accurate: The PIT value is the Cumu-  
 910 lative Distribution Function (CDF) of the  $p(z)$  evaluated at  
 911 the true redshift. A catalogue of photo- $z$  PDFs that are accu-  
 912 rate should have a flat PIT histogram (i.e., the individual  
 913 PIT values as samples from each CDF should match a Uni-  
 914 form(0,1) distribution if the CDFs are accurate). Specific  
 915 deviations from flatness indicate inaccuracy: overly broad  
 916 photo- $z$  PDFs would manifest as underrepresentation of the  
 917 lowest and highest PIT values, whereas overly narrow photo-  
 918  $z$  PDFs would manifest as over-representation of the lowest  
 919 and highest PIT values. High frequency at only PIT  $\approx 0$   
 920 and PIT  $\approx 1$  indicates the presence of catastrophic outliers  
 921 with highly inaccurate photo- $z$  PDFs where the true red-  
 922 shift is outside of the support of  $p(z)$ . Tanaka et al. (2017)  
 923 use the histogram of PIT values as a diagnostic indicator of  
 924 overall code performance, while Freeman et al. (2017) inde-  
 925 pendently define the PIT and demonstrate how its individ-  
 926 ual values may be used both to perform hypothesis testing  
 927 (via, e.g., the KS, CvM, and AD tests; see below) and to  
 928 construct quantile-quantile plots.

### 930 4.1.2 Quantile-quantile (QQ) plot

931 The quantile-quantile (QQ) plot is a graphical method for  
 932 comparing two distributions, where the quantiles of one dis-  
 933 tribution are plotted against the quantiles of the other distri-  
 934 bution (A quantile being defined by partitioning a distribu-  
 935 tion into consecutive intervals containing equal amounts of  
 936 probability, or equal numbers of objects in each interval). In  
 937 this paper we show the quantiles of the PIT values compared

<sup>17</sup> See, e. g. the LSST Data Products Definition Document, available at: <https://ls.st/dpdd>

<sup>18</sup> available at: <http://github.com/aimalz/qp>

to the quantiles of the Uniform distribution that we expect the PIT values to match if  $p(z)$  is an accurate probability distribution for all objects. 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. 14) with the true  $N(z)$ , i.e. comparing the estimated distribution of redshifts with the true redshift distribution, 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. The QQ plot contains very similar information to that shown in the PIT histogram plot, we include both forms, as visually they each convey the information in a somewhat distinct manner.

#### 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  $\mathbf{x}$ .

The CDE loss is defined as:

$$L(f, \hat{f}) = \int \int (f(z|\mathbf{x}) - \hat{f}(z|\mathbf{x}))^2 dz dP(\mathbf{x}) \quad (16)$$

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

$$\mathbb{E}_{\mathbf{X}} \left[ \int \hat{f}(z|\mathbf{X})^2 dz \right] - 2\mathbb{E}_{\mathbf{X}, Z} \left[ \hat{f}(Z|\mathbf{X}) \right] + K_f, \quad (17)$$

where 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  $f(z|\mathbf{x})$ . 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 a 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  $p(z)$  as described in Eq. 14.

### 4.2.1 Root-mean-square error (RMSE)

We employ the familiar root-mean-square error:

$$\text{RMSE} = \sqrt{\int_{-\infty}^{\infty} (\hat{f}(z) - f'(z))^2 dz}, \quad (18)$$

Though this metric does not account for the fact that the redshift distribution function is, in fact, a probability distribution, it can still be interpreted as a measure of the integrated difference between the estimated distribution and the true distribution, and it can be used to quantify the otherwise qualitative metrics.

### 4.2.2 Kolmogorov-Smirnov (KS) and related statistics

The *Kolmogorov-Smirnov statistic*  $N_{\text{KS}}$  is the maximum difference between  $F_{\text{phot}}(z)$  and  $F_{\text{spec}}(z)$ , the CDFs of the photo- $z$  and spectroscopic redshift respectively:

$$N_{\text{KS}} = \max_z (|F_{\text{phot}}(z) - F_{\text{spec}}(z)|). \quad (19)$$

The KS test quantifies the similarity between two distributions, independent of binning. A lower  $N_{\text{KS}}$  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 calculates the average of the distance squared:

$$\omega^2 = \int_{-\infty}^{+\infty} (F_{\text{meas.}}(x) - F_{\text{ideal}}(x))^2 dF_{\text{ideal}} \quad (20)$$

The AD statistic is a weighted version of the CvM statistic, making it more sensitive to the tails of the distribution:

$$A^2 = n \int_{-\infty}^{+\infty} \frac{(F_{\text{meas.}}(x) - F_{\text{ideal}}(x))^2}{F_{\text{ideal}}(x)(1 - F_{\text{ideal}}(x))} dF_{\text{ideal}} \quad (21)$$

where  $n$  is the sample size.

### 4.2.3 Moments

For the  $\hat{N}(z)$  distributions we additionally calculate the first three moments of the estimated redshift distribution 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

$$\langle z^m \rangle = \int_{-\infty}^{\infty} z^m N(z) dz. \quad (22)$$

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.

## 1028 5 RESULTS

### 1029 5.1 Ensembles of photo-z interim posteriors

1030 Fig. 1 Shows the  $p(z)$  produced by each of our eleven photo-  
 1031  $z$  codes for four example galaxies which exemplify some  
 1032 prominent cases that arise when estimating photo- $z$  PDFs: a  
 1033 narrow, unimodal redshift solution, a broader unimodal so-  
 1034 lution, a bimodal distribution, and a complex, multimodal  
 1035 distribution. The red vertical line represents the true red-  
 1036 shift of the individual galaxy, and the blue curve repre-  
 1037 sents the redshift probability. Several features are obvious  
 1038 even in these illustrative examples. ANNz2, METAPHOR,  
 1039 NN, and SKYNET all show an excess of small-scale features,  
 1040 which appear to be print-through of the underlying train-  
 1041 ing set galaxies. GPZ (in its current implementation), on  
 1042 the other hand, always produces a single Gaussian, which  
 1043 broadens to cover the multi-modal redshift solutions seen in  
 1044 other codes.

1045 As stated in Section 4,  $p(z)$  is parameterized as  $\approx 200$   
 1046 piecewise constant bins covering  $0 < z < 2$  for all eleven  
 1047 codes, giving a grid size of roughly  $\delta z = 0.01$  for each code.  
 1048 A piecewise constant grid was a natural choice for some  
 1049 photo- $z$  codes, for instance most template-based codes com-  
 1050 pute likelihoods on a fixed grid. In contrast, FlexZBoost, for  
 1051 example, can return estimates on any grid without compres-  
 1052 sion errors as its a basis expansion method where only the  
 1053 expansion coefficients need to be stored. Codes with a na-  
 1054 tive output format other than the shared piecewise constant  
 1055 binning scheme (or one that can be losslessly converted to  
 1056 it) may suffer from loss of information when converting to  
 1057 it, which could artificially favor some codes over others.

1058 Furthermore, the fidelity of photo- $z$  interim posteriors  
 1059 in this format varies with the quality of the photometry. For  
 1060 faint galaxies, this redshift resolution is sufficient to capture  
 1061 the shape of  $p(z)$  for the majority of the test sample, where  
 1062 photometric errors on the faint galaxies lead to somewhat  
 1063 broad peaks in the redshift posterior. However, as can be  
 1064 seen in e. g. the top left panel of Fig. 1, for bright galaxies  
 1065 with narrow  $p(z)$  the grid spacing of  $\delta z = 0.01$  is not suffi-  
 1066 cient to resolve the peak. This is consistent with the results  
 1067 described in Malz et al. (2018), who find that quantiles (and,  
 1068 to a lesser degree, samples) often outperform gridded  $p(z)$ ,  
 1069 particularly for bright objects and in the presence of harsher  
 1070 storage constraints. With a full 200 numbers to capture the  
 1071 information of each photo- $z$  PDF, any parametrization will  
 1072 perform adequately, but other storage parametrizations and  
 1073 limits on storage resources may be considered in future work.  
 1074 We will discuss this further in Section 6.

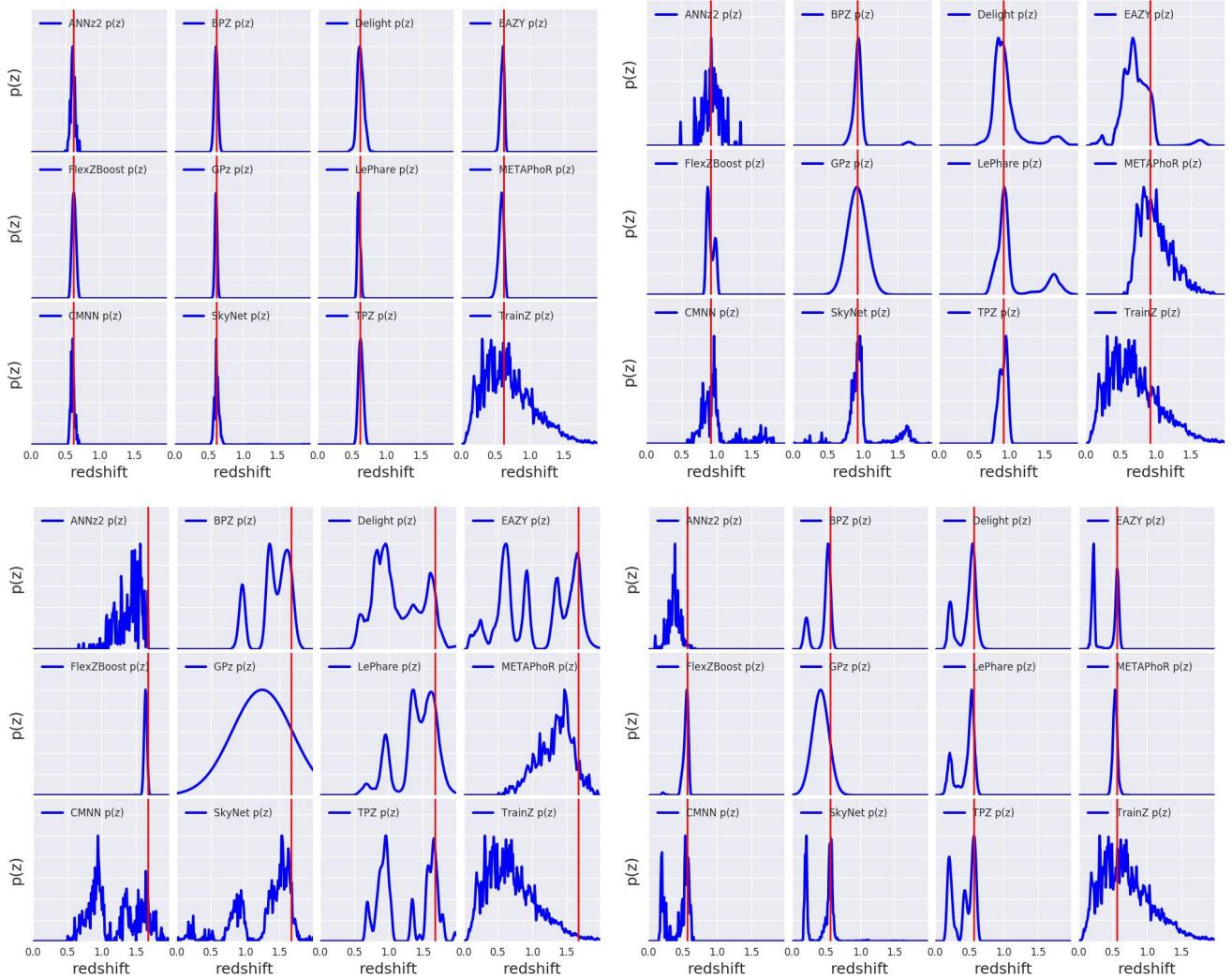
1075 Fig. 2 shows both the quantile-quantile plots (red) and  
 1076 the histogram of PIT values (blue) summarizing the results  
 1077 from each photo- $z$  code. The red line shows the measured  
 1078 quantiles, while the black diagonal represents the ideal QQ  
 1079 values if the distribution were perfectly reproduced. A sec-  
 1080 ond panel below the main panel for each code shows the dif-  
 1081 ference between  $Q_{data}$  and  $Q_{theory}$ , i. e. the departure from  
 1082 the diagonal, for clarity. Biases and trends in whether the  
 1083 average width of the  $p(z)$  values being over/under-predicted  
 1084 are evident. An overall bias where the predicted redshift  
 1085 is systematically low manifests as the measured QQ value  
 1086 falling above the diagonal, as is the case for BPZ and EAZY,  
 1087 while a systematic overprediction shows up as the measured  
 1088 QQ value falling below the diagonal, as seen in TPZ. In

1089 terms of PIT histograms, a systematic underprediction of  
 1090 redshift corresponds to fewer PIT values at  $PIT < 0.5$  and  
 1091 more at  $PIT > 0.5$ , while a systematic overprediction will  
 1092 show the opposite.

1093 Examination of the PIT histograms and QQ plots shows  
 1094 that there are fairly generic issues with the width of  $p(z)$  un-  
 1095 certainties: DELIGHT, NN, SKYNET and TPZ all show a PIT  
 1096 histogram with an dearth of low values and an excess of  
 1097 high values, signs that, on average, their  $p(z)$  are more  
 1098 broad than the true distribution of redshifts. METAPHOR  
 1099 shows the opposite trend, indicating the  $p(z)$  are more  
 1100 narrow than the distributions given by the true redshifts.  
 1101 In all of these code cases there is a free parameter or band-  
 1102 width that can be used to tune uncertainties. The sensitivity  
 1103 of multiple codes to this bandwidth choice emphasizes the  
 1104 fact that great care must be taken in setting user-defined  
 1105 parameters in photo- $z$  codes, even in the presence of rep-  
 1106 resentative training/validation data. for FLEXZBOOST the  
 1107 “sharpening” parameter (described in Section 3.2.4) plays  
 1108 a key role in improving the results, resulting in a QQ plot  
 1109 that is very nearly diagonal. A similar sharpening procedure  
 1110 could be beneficial for several codes. Interestingly, the three  
 1111 purely template-based codes, BPZ, EAZY, and LEPHARE,  
 1112 show relatively well behaved  $p(z)$  statistics (albeit with some  
 1113 bias), which may indicate that the likelihood estimation with  
 1114 representative templates is accurately capturing the uncer-  
 1115 tainties on individual redshifts.

1116 The ideal PIT histogram would follow the black dashed  
 1117 line, representing a uniform distribution of PIT values,  
 1118 equivalent to the diagonal line in the QQ plot. Overly broad  
 1119  $p(z)$  values show up as an excess of PIT values near 0.5  
 1120 and a dearth of values at the edges, while overly narrow  
 1121  $p(z)$  will have an excess at the edges and will be missing  
 1122 values at the centre. Another feature evident in the PIT  
 1123 histograms is the number of “catastrophic outlier” values  
 1124 where the true redshift falls outside of the non-zero support  
 1125 of  $p(z)$ , corresponding to  $PIT = 0.0$  or  $1.0$  is more apparent  
 1126 than in the QQ plots. Following Kodra & Newman (in prep.)  
 1127 we define  $f_0$  as the fraction of objects with  $PIT < 0.0001$   
 1128 or  $PIT > 0.9999$ . Table 2 lists these fractions for each of  
 1129 the codes. For a proper Uniform distribution we expect a  
 1130 value of 0.0002. Several codes show a marked excess, with  
 1131 ANNz2, FLEXZBOOST, LEPHARE, AND METAPHOR with  
 1132  $f_0 > 0.02$ , indicating a sizeable number of catastrophic red-  
 1133 shift solutions where the true redshift is not covered by the  
 1134 extent of  $p(z)$ . For METAPHOR this may be partially due  
 1135 to an overall underprediction of the  $p(z)$  width, however this  
 1136 is not the case for the other codes. LEPHARE is a particular  
 1137 outlier with nearly 5 per cent of objects outside of  $p(z)$  sup-  
 1138 port. Further study will be necessary to determine what is  
 1139 causing these misclassifications for LEPHARE. As expected,  
 1140 and by design, TRAINZ has the proper fraction of outliers  
 1141 for the  $f_0$  statistic.

1142 Fig. 3 shows comparative metric values for the quantita-  
 1143 tive Kolmogorov-Smirnoff (KS), Cramer-Von Mises (CvM),  
 1144 and Anderson Darling (AD) test statistics for each of the  
 1145 codes based on comparing the distribution of their PIT val-  
 1146 ues to the expected uniform distribution over the interval  
 1147  $[0,1]$ . The individual values of the statistic are not as impor-  
 1148 tant as the comparative score between the different codes.  
 1149 The AD test statistic diverges for values that include the  
 1150 extrema, and thus is calculated by excluding the edges of the



**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.

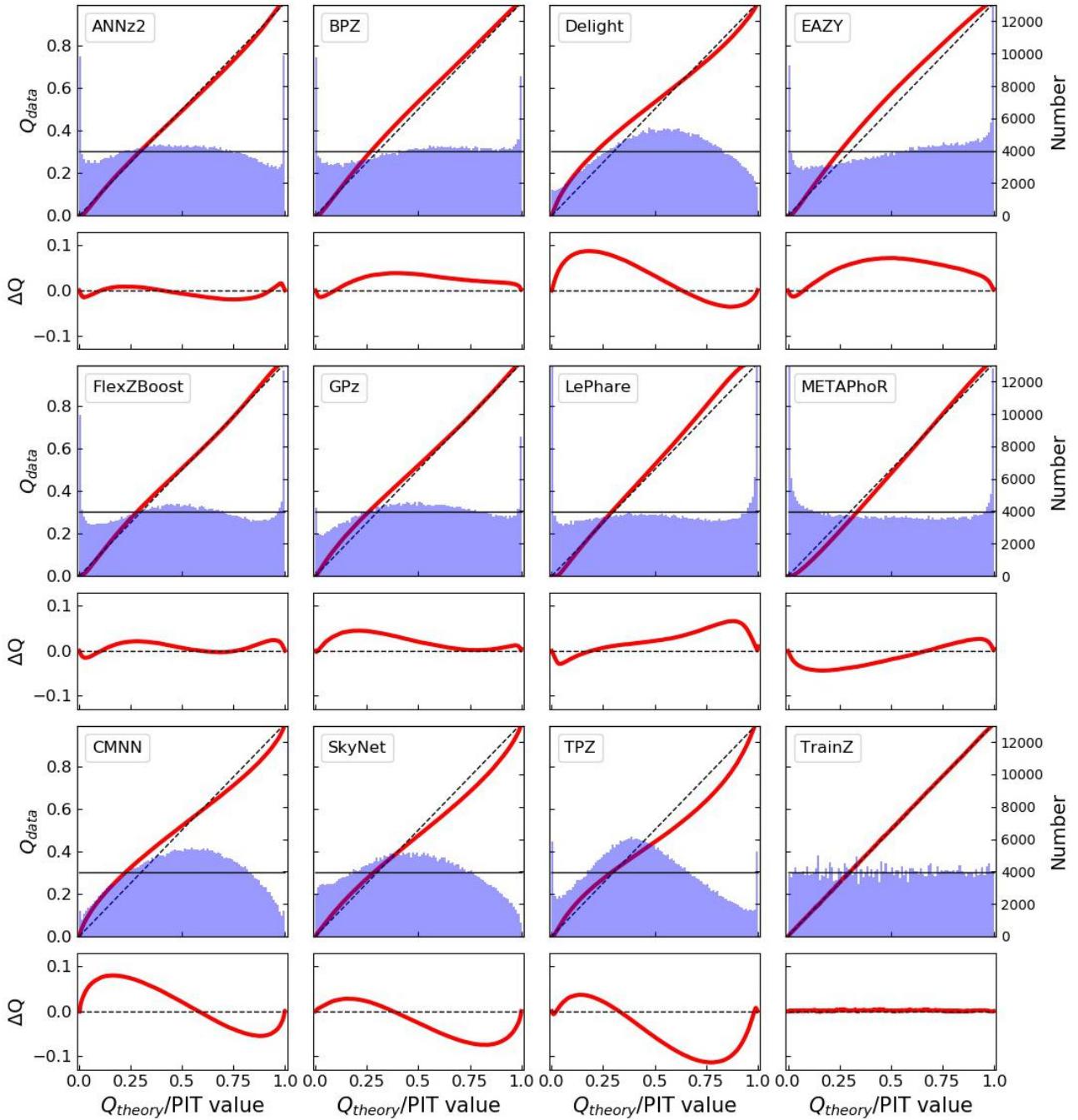
1151 distribution. We calculate the AD statistic over the range 1169  
 1152 of PIT values  $v = [0.01, 0.99]$ . ANNz2 and FLEXZBOOST 1170  
 1153 score very well for the PIT metrics. METAPHOR and LEP- 1171  
 1154 HARE score very well in the PIT AD statistic, but both have 1172  
 1155 a large number of catastrophic outliers, resulting in higher 1173  
 1156 KS and CvM scores.

1157 Given the near-perfect training data, examining the individual 1174  
 1158 codes for explanations for departures from the expected 1175  
 1159 behaviour will be instructive in avoiding similar problems 1176  
 1160 in future tests. ANNz2 performs quite well in  $p(z)$  1177  
 1161 based metrics. In the specific implementation employed in 1178  
 1162 this paper, the final  $p(z)$  is a weighted average of five neural- 1179  
 1163 nets. During the training process ANNz2 compares the per- 1180  
 1164 centiles of the redshift training sample against the CDFs of 1181  
 1165 the  $p(z)$  sample. Distributions that more closely match are 1182  
 1166 given extra weight, and the final weights are designed to 1183  
 1167 produce accurate percentiles. Given that our metrics are fo- 1184  
 1168 cused on the percentile distributions, it is unsurprising that 1185

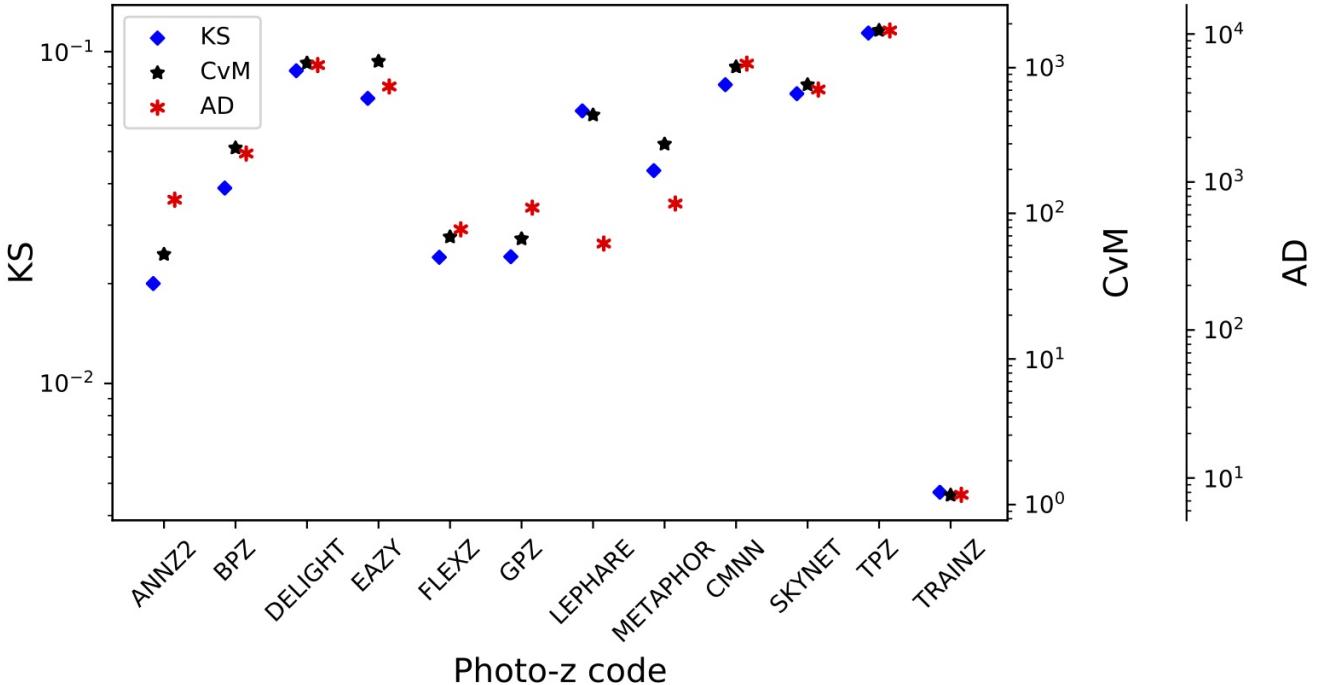
ANNz2 performs well in the given metrics. The discreteness in the individual  $p(z)$  estimated by ANNz2 can be attributed to the fact that the code was run as a classifier, assigning weights to discrete bins of redshift. While multiple bins may receive weight, the bins themselves will still be discretized, 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 smoothed via kernel density estimation (KDE), with a bandwidth chosen via Scott's rule (Scott 1992). While Scott's rule is used to dis-



**Figure 2.** Summary plots for all eleven 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-Smirnov (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, though the AD statistic truncates the extrema of the distribution and can have disparate values compared to KS and CvM.

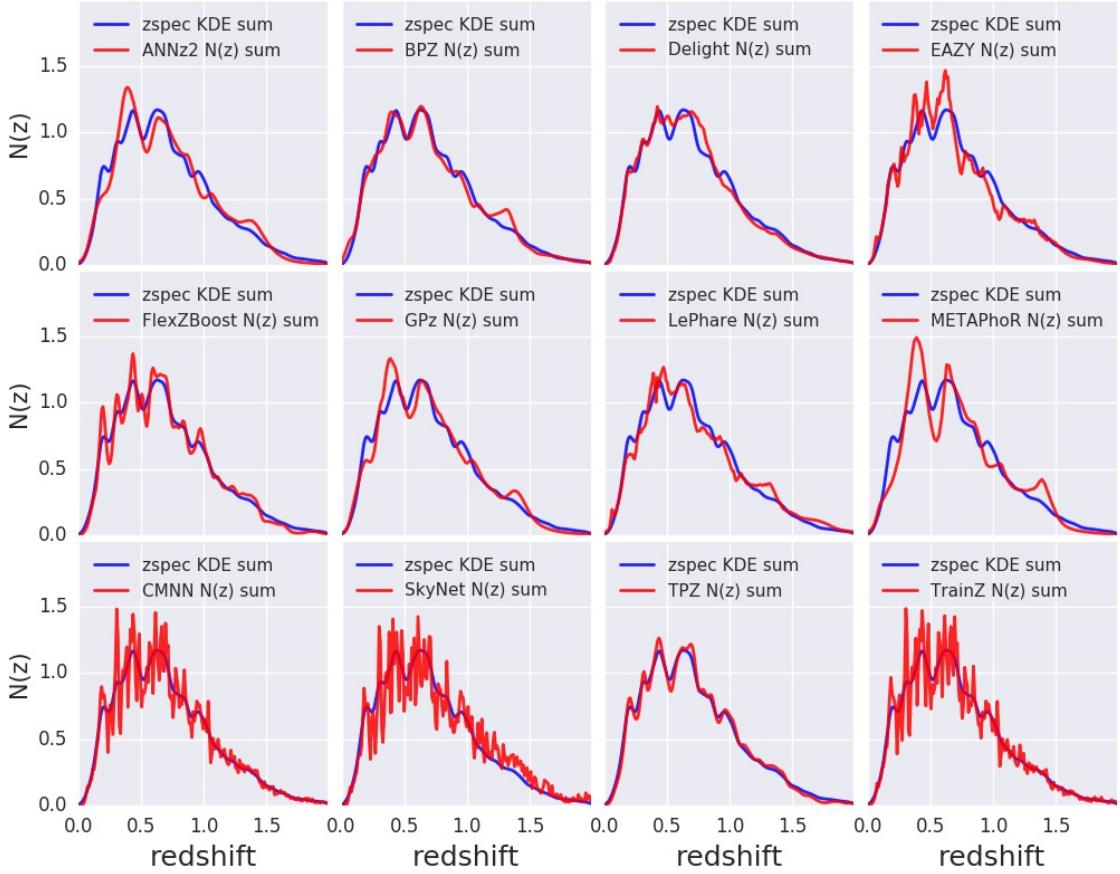
**Table 2.** The fraction of “catastrophic outlier” PIT values. We expect a value of 0.0002 for a proper Uniform distribution. An excess over this small value indicates true redshifts that fall outside the non-zero support of the  $p(z)$ .

Photo-z Code	fraction $\text{PIT} < 10^{-4}$ 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

play  $N'(z)$  in the figure, all quantitative statistics are computed via the empirical CDF, and are thus unaffected by bandwidth/smoothing choice. Several of the 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 angstrom break passing through the gap between the  $z$  and  $y$  filters. This feature is one of the most prominent in individual galaxy  $p(z)$ , and is readily seen in the point-estimate plots shown in Fig. A1 and described in the Appendix. Sev-

eral of the machine learning based codes appear to be over-trained, adding excess galaxy probability to the redshift peaks and missing probability in the troughs. Given 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. A more extensive training/validation set might allow for a better choice of smoothing parameters in individual codes that would avoid such overtraining.

As with the  $p(z)$  values in Figure 2, different levels of substructure are obvious for the different codes. While Scott’s rule provides a relatively good general smoothing scale to represent the true  $N'(z)$ , there are smaller scale fluctuations: while FLEXZBOOST and CMNN appear somewhat discrepant in Fig. 4, they are actually the two most accurate in terms of their quantitative measurements. Interestingly, while ANNz2 shows an abundance of small scale structure in individual  $p(z)$  measurements (see Fig. 1), the summed  $\hat{N}(z)$  is rather smooth, where the small scale features average out. This is not the case for the two other codes that show an abundance of substructure in their individual  $p(z)$ : both CMNN and SKYNET show small scale features both in  $p(z)$  and  $\hat{N}(z)$ . For CMNN the  $p(z)$  are simply a weighted histogram of all spectroscopic training galaxies in nearby colour space with no smoothing applied, so the substructure is not unexpected. The PIT histogram and shape of the QQ plot in Figure 2 show that CMNN is producing  $p(z)$  that are overly broad, additional smoothing of the  $p(z)$  would exacerbate this problem. While the  $\hat{N}(z)$  plot shows more small scale features than other codes, these features are actually representative of real structure in the



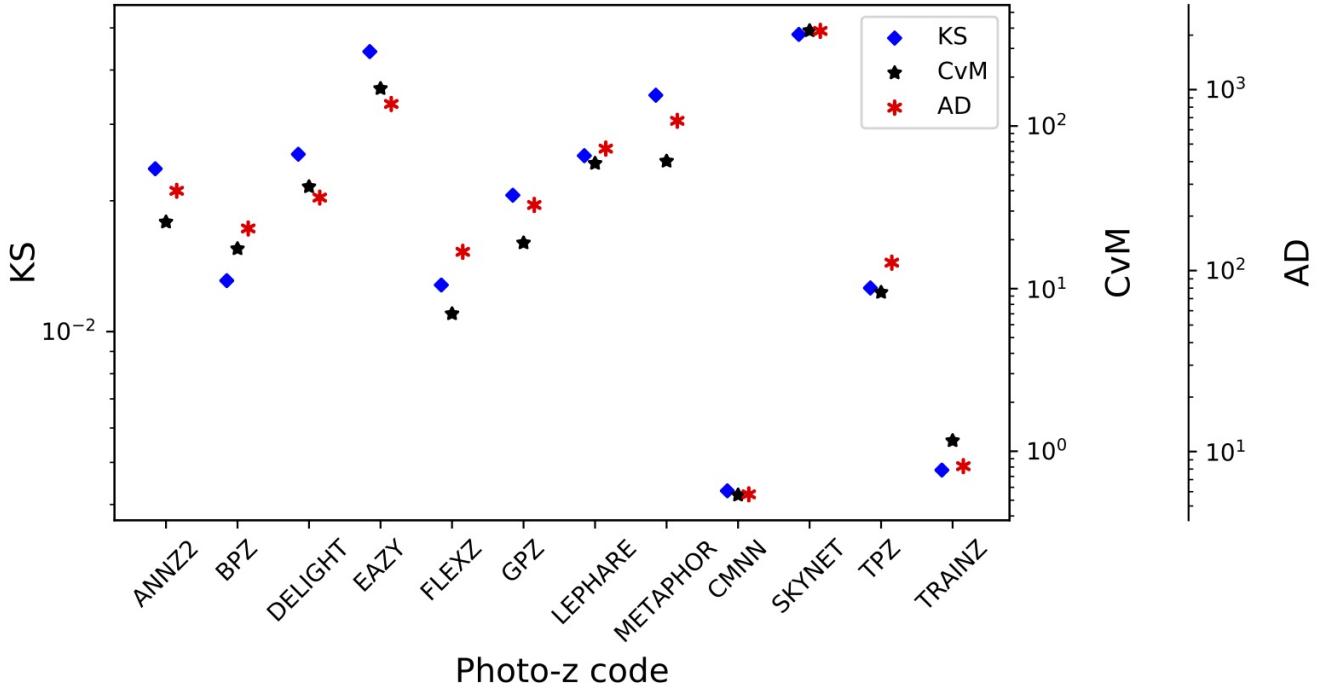
**Figure 4.** The stacked  $p(z)$  produced by each photo- $z$  code ( $\hat{N}(z)$ , red) compared to the spectroscopic redshift distribution ( $N'(z)$ , blue). Varying levels of small-scale structure are seen in the codes.  $N'(z)$  is smoothed using a single bandwidth chosen via Scott's rule for all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

true  $N'(z)$ , as evidenced by the very good metric scores for CMNN. SKYNET  $p(z)$  were also not smoothed: 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. In addition to excess substructure, SKYNET shows an obvious redshift bias, evident both visually in Figure 4 and in the first moment of  $N(z)$  listed in Table 5, 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. Also unlike 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.

It is also of note that all three template-based codes show an excess in their stacked  $p(z)$  at  $z \sim 1.3 - 1.4$ . This redshift range corresponds to the wavelengths where the 4000 Angstrom break is passing between the borders of the  $z$  and  $y$  filters. This strong break entering the gap between the two reddest filters can cause problems with redshift estimation of individual galaxies, as can be seen in the



**Figure 5.** A visual representation of the Kolmogorov-Smirnov (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.

point-estimate photo- $z$ 's shown in Figure A1. This 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.

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. Empirically, we have found that PIT RMSE is not as closely correlated to CDE loss as it is to the  $N(z)$  statistics. 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 gives the root-mean-square-error (RMSE) statistics for both the PIT and  $N(z)$  estimators. The PIT value calculates the RMSE between the quantiles shown in the QQ plot in Figure 2 and the diagonal, while the  $N(z)$

calculates the RMSE between the cumulative distribution of the stacked  $\hat{N}(z)$  and the true redshift distribution  $N'(z)$ .

Table 5 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. 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.

1324 **5.3 Interpretation of metrics**

1325 Samples from accurate photo- $z$  posteriors should reproduce  
 1326 the space of  $p(z, \text{data})$ . However, it is difficult to test this re-  
 1327 construction given our data set, as the galaxy distributions  
 1328 arise from mock objects pasted on to an underlying dark  
 1329 matter halo catalogue with properties designed to match  
 1330 empirical relations, rather than being drawn from statisti-  
 1331 cal distributions in redshift. In previous sections we have  
 1332 mentioned that optimizing for a specific metric does not  
 1333 guarantee good performance on other metrics, nor is there  
 1334 any guarantee that good performance by our metrics cor-  
 1335 responds to *accurate* photo- $z$  posteriors. In other words, we  
 1336 can construct photo- $z$  estimators that provide good coverage  
 1337 in many of our tests, but which have very little predictive  
 1338 power.

1339 The TRAINZ estimator, which assigns every galaxy a  
 1340  $p(z)$  equal to  $N(z)$  of the training set as described in Sec-  
 1341 tion 3.3, is introduced as a “null test” to demonstrate this  
 1342 point via *reductio ad absurdum*. TRAINZ outperforms all  
 1343 codes on the PIT-based metrics, and all but one code on  
 1344 the  $N(z)$  based statistics. Because our training set is per-  
 1345 fectly representative of the test set,  $N(z)$  should be identical  
 1346 for both sets down to statistical noise.

1347 The CDE loss and point estimate metrics, however, suc-  
 1348 cessfully identify problems with TRAINZ. As shown in Ap-  
 1349 pendix A, TRAINZ has identical  $ZPEAK$  and  $ZWEIGHT$   
 1350 values for every galaxy, and thus the photo- $zs$  are constant  
 1351 as a function of spec- $zs$ , i.e. a horizontal line at the mode  
 1352 and mean of the training set distribution respectively. The  
 1353 explicit dependence on the *individual posteriors in the cal-*  
 1354 *culation of the CDE loss, described in Section 4.1.3, dis-*  
 1355 *tinguishes this metric from the other  $p(z)$  metrics that test*  
 1356 *the overall ensemble of  $p(z)$  distributions. With a represen-*  
 1357 *tative training set, TRAINZ will score well on the ensemble*  
 1358 *metrics, but fails miserably for metrics tied to individual red-*  
 1359 *shifts. We note that many of the ensemble-based metrics are*  
 1360 *prominent in the photo- $z$  literature despite their inability to*  
 1361 *identify problems such as those exemplified by TRAINZ.*

1362 *In summary, context is crucial to interpreting metrics*  
 1363 *and defending against the likes of TRAINZ. The best photo- $z$*   
 1364 *method is the one that most effectively achieves our science*  
 1365 *goals, not the one that performs best on a metric that does*  
 1366 *not accurately reflect those goals. In the absence of clear*  
 1367 *goals or the information necessary for a principled metric*  
 1368 *definition, we must think carefully before choosing a single*  
 1369 *metric*

1370 **6 DISCUSSION**

1371 *In this paper we presented results evaluating the photomet-*  
 1372 *ric redshift PDF computation for eleven photo- $z$  codes. As*  
 1373 *discussed in Section 4 the  $p(z)$  should accurately reflect the*  
 1374 *relative likelihood as a function of redshift for each galaxy.*  
 1375 *All codes were provided a set of representative training data*  
 1376 *and tested on an idealized set of model galaxies with high*  
 1377 *signal-to-noise and photometry with no confounding effects*  
 1378 *due to blending, instrumental effects, the night sky, etc...  
 1379 included. The goal was not to determine a “best” photo- $z$*   
 1380 *code: in many ways, this was a baseline test of a “best case*  
 1381 *scenario” to predict the expected photo- $z$  performance if a*

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. Root-Mean-Square-Error (RMSE) statistics for the eleven photo- $z$  codes for both PIT and  $\hat{N}(z)$  distributions.

Photo- $z$ Code	Root-Mean-Square-Error (RMSE) statistics	
	PIT RMSE	$N(z)$ RMSE
ANNz2	0.019	0.0054
BPZ	0.032	0.0050
DELIGHT	0.111	0.0056
EAZY	0.054	0.0102
FLEXZBOOST	0.021	0.0022
GPz	0.027	0.0042
LEPHARE	0.028	0.0062
METAPHOR	0.064	0.0081
CMNN	0.108	0.0009
SKYNET	0.054	0.0144
TPZ	0.082	0.0031
TRAINZ	0.0025	0.0013

Table 5. 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

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 in a problem with either/both of the specific code implementation or the underlying algorithm. In order to meet the stringent LSST requirements on photo-z performance, identifying and correcting such problems is an important first step before tackling more realistic data in future challenges. Most of the codes tested performed well, however, several did not meet the stringent goals that have been laid out for LSST photometric redshift performance. This is a cause for concern, given the idealized conditions, and the individual code responses will be studied in detail moving forward. 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  $\bar{N}(z)$  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  $\bar{N}(z)$  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.

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. [are there any references for this? I remember Gary Bernstein talking about this at a photo-z workshop in Japan, but I don't know that it was published. I believe Michael Troxel has discussed this as well.] We also plan to propagate the uncertainties measured in a set of fiducial tomographic redshift bins in order to estimate impact on cosmological parameter estimation.

In future papers we will 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. 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. However, due to a combination of factors due to both the galaxy SEDs and limitations of spectrographic instruments, redshift samples are known to be systematically incomplete, where certain galaxy types and redshift intervals fail to yield a redshift even at the longest integration times on current and near-future instruments. The more representative the training data, the better the performance of photo-z algorithms will be. Current and upcoming surveys are putting in 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. One major focus of an upcoming LSST Dark Energy Science Collaboration Photo-z Working Group data challenge is 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.

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. Object blending will be a major area of investigation, as the mixing of flux from multiple

## 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

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. Inclusion of differing observing conditions (seeing, clouds, variations in filter curves, Galactic dust, ...), as well as models for instrumental and system effects, sky masks, will all impact object photometry, and will be explored in the upcoming data challenge and their impacts described in upcoming papers. All underlying SEDs were parameterized as a weighted combination of five basis SEDs, with no additional accounting for host galaxy dust obscuration beyond what was encoded in the basis templates. This, in effect, limited the simulation to a very simple model of internal obscuration. Future simulations will include a more complicated and realistic treatment of host galaxy dust.

The underlying simulation used in this work was based on a light-cone constructed 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.

## 7 CONCLUSION

In this study we have not accounted for the presence of Active Galactic Nuclei (AGN) contributions to galaxy fluxes. In some cases, AGN will be easily identified from the colors and morphologies, i.e. the case of the brightest quasars where the nuclear activity outshines the host galaxy, and numerous studies have utilized color selection to create large samples of quasars (e.g. Richards *et al.* 2006; Maddox *et al.* 2008; Richards *et al.* 2015). In current deep fields, similar in depth to what we expect from LSST, variability information and multi-wavelength data have been critical to not only identify AGN dominated galaxies, but also obtain more accurate photometric redshifts (e.g. Salvato *et al.* 2011).

In addition to AGN dominated galaxies, those with lower levels of nuclear activity present 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. For LSST, we will need to identify and obtain accurate photometric redshifts of all types of AGN for a range of science goals, whether it is to eliminate such objects from cosmology

experiments, or to use them with confidence, all the way through to understanding galaxy evolution and the role that AGN may play in influencing galaxy properties over cosmic time.

A promising route to classifying and obtaining accurate photometric redshifts for the AGN population is by combining machine learning with template-fitting techniques, as has recently been demonstrated by Duncan *et al.* (2018) for radio-selected AGN. This is because AGN are relatively easy to obtain spectroscopic redshifts for over all redshifts due to the strong emission lines that they exhibit, allowing very good training sets for machine learning algorithms to use. Whereas for those galaxies where the AGN is sub-dominant the galaxy templates are still adequate for obtaining reasonable photometric redshifts.

In addition to these improvements, the DESC Photo-z group plans to look at all potential methods to combine the results from multiple photo-z codes to improve  $p(z)$  accuracy, similar to the work presented in Dahlen *et al.* (2013); Carrasco Kind & Brunner (2014); Duncan *et al.* (2018). Taking advantage of multiple algorithms that use observables in slightly different ways has shown promise, however we must be very conscious of whether a potential combination properly treats the covariance between the methods, given that they are estimating quantities based on the same underlying observables. Several science cases wish to estimate physical quantities along with redshift, for example galaxy stellar mass and star formation rate. Proper joint estimation of redshift and physical quantities requires an in depth understanding of galaxy evolution, and progress on accurate bivariate redshift probability distributions will go hand in hand with progress on understanding galaxies themselves. Parameterization and storage of a complex 2-dimensional probability surface for potentially billions of galaxies (or even subsets of hundreds of thousand of particular interest) pose a potential challenge. These issues will be examined in another future paper.

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 series of future 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)

1622 J.Y.H. Soo: Ran ANNz2 and Delight, updated abstract, 1684  
 1623 edited sections 1 through 6, added tables in Methods 1685  
 1624 and Results, updated references.bib and added references 1686  
 1625 throughout the paper 1687  
 1626 M. Brescia: main ideator of METAPHOR and of MLPQNA; 1688  
 1627 modification of METAPHOR pipeline to fit the LSST data 1689  
 1628 structure and requirements 1690  
 1629 S. Cavaudi: Contributed to choice and test of metrics, ran 1691  
 1630 METAPHOR, minor text editing 1692  
 1631 G. Longo: Scientific advise, test and validation of the 1693  
 1632 modified METAPHOR pipeline, text of the METAPHOR 1694  
 1633 section  
 1634 I.A. Almosallam: vetted the early versions of the data set  
 1635 and ran many photo-z codes on it, applied GPz to the final  
 1636 version and wrote the GPz subsection  
 1637 M.L. Graham: Ran the colour-matched nearest-neighbours  
 1638 photo-z code on the Buzzard catalog and wrote the relevant  
 1639 piece of Section 2; participated in discussions of the analy- 1695  
 1640 sis.  
 1641 A.J. Connolly: Developed the colour-matched nearest- 1696  
 1642 neighbours photo-z code; participated in discussions of the 1697  
 1643 analysis.  
 1644 E. Nourbakhsh: Ran and optimized TPZ code on the 1699  
 1645 Buzzard catalog and wrote a subsection of Section 2 for that 1700  
 1646 J. Cohen-Tanugi: contributed to running code, analysis 1701  
 1647 discussion, and editing, reviewing the paper 1702  
 1648 H. Tranin: contributed to providing SkyNet results and 1703  
 1649 writing the relevant section 1704  
 1650 P.E. Freeman: Contributed to choice of CDE metrics and 1705  
 1651 to implementation of FlexZBoost 1706  
 1652 K. Iyer: assisted in writing metric functions used to evaluate 1707  
 1653 codes 1708  
 1654 J.B. Kalmbach: Worked on preparing the figures for the 1709  
 1655 paper.  
 1656 E. Kovacs: Ran simulations, discussed data format and 1710  
 1657 properties for SEDs, dust, and ELG corrections 1711  
 1658 A.B. Lee: Co-developed FlexZBoost and the CDE loss statis- 1712  
 1659 tic, wrote text on the work, and supervised the development 1713  
 1660 of FlexZBoost software packages 1714  
 1661 C. Morrison: Managerial support; Discussions with authors 1715  
 1662 regarding metrics and style; Some coding contribution to 1716  
 1663 metric computation.  
 1664 J. Newman: Contributions to overall strategy, design of 1717  
 1665 metrics, and supervision of work done by Rongpu Zhou 1718  
 1666 E. Nuss: contributed to running code, analysis discussion, 1721  
 1667 and editing, reviewing the paper 1722  
 1668 T. Pospisil: Co-developed FlexZBoost software and CDE 1723  
 1669 loss calculation code 1724  
 1670 M.J. Jarvis: Contributed text on AGN to Discussion section 1725  
 1671 and portions of GPz work 1726  
 1672 R. Izbicki: Co-developed FlexZBoost and the CDE loss 1727  
 1673 statistic, and wrote software for FlexZBoost 1728

1675 The authors would like to thank their LSST-DESC pub- 1730  
 1676 lication review committee.  
 1677

1678 AIM is advised by David W. Hogg and was supported by 1732  
 1679 National Science Foundation grant AST-1517237.  
 1733

1680 The DESC acknowledges ongoing support from the In- 1734  
 1681 stitut National de Physique Nucléaire et de Physique des 1735  
 1682 Particules in France; the Science & Technology Facilities 1736  
 1683 Council in the United Kingdom; and the Department of En- 1737  
 1684 ergy, the National Science Foundation, and the LSST Cor- 1738

1685 poration in the United States. DESC uses resources of the  
 1686 IN2P3 Computing Center (CC-IN2P3-Lyon/Villeurbanne -  
 1687 France) funded by the Centre National de la Recherche Sci-  
 1688 entifique; the National Energy Research Scientific Comput-  
 1689 ing Center, a DOE Office of Science User Facility supported  
 1690 by the Office of Science of the U.S. Department of Energy  
 1691 under Contract No. DE-AC02-05CH11231; STFC DiRAC  
 1692 HPC Facilities, funded by UK BIS National E-infrastructure  
 1693 capital grants; and the UK particle physics grid, supported  
 1694 by the GridPP Collaboration. This work was performed in  
 1695 part under DOE Contract DE-AC02-76SF00515.

## APPENDIX A: POINT ESTIMATE PHOTOMETRIC REDSHIFTS

While we do not recommend the use of single point estimates of redshift for most science applications, 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 comparisons with previous 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,  $z_{\text{PEAK}}$  and  $z_{\text{WEIGHT}}$ .  $z_{\text{PEAK}}$  is simply the maximum value attained for each galaxy  $p(z)$ , the mode of the probability distribution.  $z_{\text{WEIGHT}}$  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  $z_{\text{PEAK}}$  value. The main peak is defined by subtracting  $0.05 \times z_{\text{PEAK}}$  from  $p(z)$  and identifying the roots to isolate the peak containing  $z_{\text{PEAK}}$ ,  $z_{\text{WEIGHT}}$  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.

## 1739 A1 Point Estimate Metrics

1740 We calculate the commonly used point estimate metrics of  
 1741 the overall photo-z scatter ( $\sigma_z$ , the standard deviation of  
 1742 the photo-z residuals), bias, and “catastrophic outlier rate”.  
 1743 Specifically, we calculate the metrics as follows: we define  $e_z$   
 1744 as

$$1745 e_z = \frac{z_p - z_s}{1 + z_s} \quad (\text{A1})$$

1746 where  $z_p$  is the point estimate and  $z_s$  is the true redshift. In  
 1747 practice, because the standard deviation calculation is quite  
 1748 sensitive to the outliers, we define the photo-z scatter,  $\sigma$  in  
 1749 terms of the Interquartile Range (IQR), the difference be-  
 1750 tween the 75th and 25th percentiles of the  $e_z$  distribution.  
 1751 In order to match the usual meaning of a  $1\sigma$  interval, we  
 1752 scale the IQR and define  $\sigma_{IQR} = IQR/1.349$ , as there is  
 1753 a factor of 1.349 difference between the IQR and the stan-  
 1754 dard deviation of a Normal distribution. While many other  
 1755 studies define the bias based on the mean offset between true  
 1756 and estimated redshift, in this study we define the bias as  
 1757 the median value of  $e_z$  for the sample. We use median as  
 1758 it is, once again, less sensitive to outliers than the mean.  
 1759 The catastrophic outlier fraction is defined as the fraction  
 1760 of galaxies with  $e_z$  greater than the larger of  $3\sigma_{IQR}$  or 0.06,  
 1761 i.e.  $3\sigma$  outliers with a floor of  $\sigma_{IQR}=0.02$ . For reference, the  
 1762 goals stated in Section 3.8 of the LSST Science Book (Abell  
 1763 et al. 2009) for photo-z performance in these metrics, as-  
 1764 suming perfect training knowledge (as we are testing in this  
 1765 paper) are:

- 1766 • RMS scatter <  $0.02(1+z)$
- 1767 • bias < 0.003
- 1768 • catastrophic outlier rate < 10%

1769 These definitions are similar, but not exactly the same, as  
 1770 the  $\sigma_{IQR}$  and median bias calculated here, but are similar  
 1771 enough for qualitative comparisons to the LSST goals.

1772 Fig. A1 shows the point estimates for both  $z_{\text{PEAK}}$  and  
 1773  $z_{\text{WEIGHT}}$ . Point density is shown with mixed contours to  
 1774 emphasize that most of the galaxies do fall close to the  
 1775  $z_{\text{phot}} = z_{\text{spec}}$  line, while blue points show differing char-  
 1776 acteristics of the outlier populations. The red dashed lines  
 1777 indicated the cutoff for catastrophic outliers, defined as:  
 1778  $\max(0.06, 3\sigma_{IQR})$ . As with the full  $p(z)$  results, a variety  
 1779 of behaviours are evident in the different codes. Table A1  
 1780 lists the scatter, bias, and catastrophic outlier fractions for  
 1781 the codes. The performance of the codes for point metrics is  
 1782 highly correlated with performance on  $p(z)$  based tests, which  
 1783 is to be expected, given that the point-estimates were derived  
 1784 from the  $p(z)$ . Some discretization is evident in  $z_{\text{PEAK}}$ , par-  
 1785 ticularly for SKYNET, due to the finite grid spacing of the  
 1786 reported  $p(z)$ . These discreteness effects are mitigated by the  
 1787 weighting of  $z_{\text{WEIGHT}}$ , resulting in a smoother distribution  
 1788 of redshift estimates. Several features perpendicular to the  
 1789 main  $z_{\text{phot}} = z_{\text{spec}}$  line are evident. These features are due  
 1790 to the 4000 angstrom break passing through the gaps between  
 1791 adjacent LSST filters. These features are most prominent in  
 1792 template-based codes, but appear to some degree in all codes  
 1793 tested.

1794 In even the best performing codes, there are visible occu-  
 1795 pied regions away from the  $z_{\text{phot}} = z_{\text{spec}}$  line, corresponding  
 1796 to degenerate redshift solutions for certain LSST magnitudes

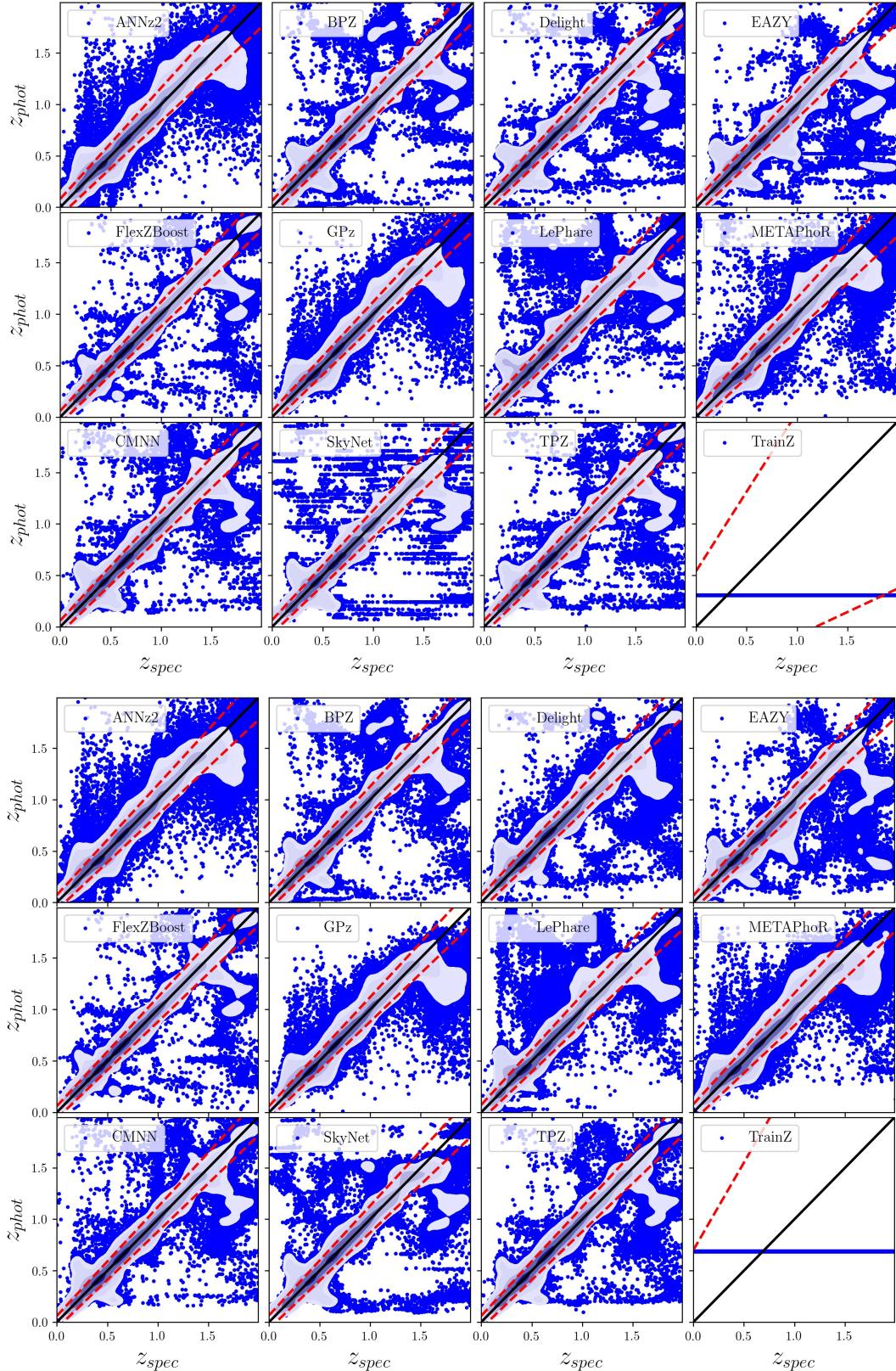
1797 and colors. While use of the full information available via  
 1798  $p(z)$  mitigates their impact, a full understanding of the out-  
 1799 lier population is critical for LSST science, particularly in  
 1800 tomographic applications

1801 Finally, we note that all eleven codes are at or near the  
 1802 goals for point-estimates as outlined in the LSST Science  
 1803 Requirements Document<sup>19</sup> and Graham et al. (2018). This  
 1804 is to be expected, given that the requirements were designed  
 1805 such that a point estimate photo-z would meet these require-  
 1806 ments for perfect training data to a depth of  $i < 25$ . But, it  
 1807 is still an encouraging sign, given an updated mock galaxy  
 1808 simulation and the expanded set of photo-z codes tested.

## REFERENCES

- Abbott T., et al., 2005, preprint ([arXiv:astro-ph/0510346](https://arxiv.org/abs/astro-ph/0510346))  
 Abell P. A., et al., 2009, preprint ([arXiv:0912.0201](https://arxiv.org/abs/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  
 Biviano A., et al., 2013, *A&A*, **558**, A1  
 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., et al., 2016, *Phys. Rev. D*, **94**, 042005  
 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, Califor-  
 nia, U.S.A  
 Brescia M., Cavuoti S., D'Abrusco R., Longo G., Mercurio A.,  
 2013, *ApJ*, 772  
 Brescia M., Cavuoti S., Longo G., De Stefano V., 2014, *A&A*,  
 568  
 Brescia M., Cavuoti S., Amaro V., Riccio G., Angora G., Vel-  
 lucchi C., Longo G., 2018, preprint, ([arXiv:1802.07683](https://arxiv.org/abs/1802.07683))  
 Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, **432**, 1483  
 Carrasco Kind M., Brunner R. J., 2014, *MNRAS*, **442**, 3380  
 Cavuoti S., Brescia M., De Stefano V., Longo G., 2015, *Exp.  
 Astron.*, **39**, 45  
 Cavuoti S., Amaro V., Brescia M., Vellucci C., Tortora C.,  
 Longo G., 2017a, *MNRAS*, **465**, 1959  
 Cavuoti S., et al., 2017b, *MNRAS*, **466**, 2039  
 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>  
 Dahlen T., et al., 2013, *ApJ*, **775**, 93  
 Dawson W. A., Schneider M. D., Tyson J. A., Jee M. J., 2016,  
*ApJ*, **816**, 11  
 Duncan K. J., Jarvis M. J., Brown M. J. I., Röttgering H. J. A.,  
 2018, *Monthly Notices of the Royal Astronomical Society*,  
 p. 940

1795 <sup>19</sup> available at: <http://ls.st/srd>



**Figure A1.** Point estimate photo-z's derived from the posteriors. Top panel shows  $z_{\text{PEAK}}$ , while bottom panel shows  $z_{\text{WEIGHT}}$ . 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

Photo-z Code	ZPEAK			ZW EIGHT		
	$\frac{\sigma_{IQR}}{(1+z)}$	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

- 1857 Fernández-Soto A., Lanzetta K. M., Yahil A., 1999, *ApJ*, **513**, 1903  
1858      34  
1859 Firth A. E., Lahav O., Somerville R. S., 2003, *MNRAS*, **339**, 1905  
1860      1195  
1861 Freeman P. E., Izbicki R., Lee A. B., 2017, *MNRAS*, **468**, 4556  
1862 Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, *MNRAS*, 1908  
1863      441, 1741  
1864 Graham M. L., Connolly A. J., Ivezić Ž., Schmidt S. J., Jones 1910  
1865      R. L., Jurić M., Daniel S. F., Yoachim P., 2018, *AJ*, **155**, 1  
1866 Green J., et al., 2012, preprint (arXiv:1208.4012), 1912  
1867 Hildebrandt H., et al., 2010, *A&A*, **523**, A31  
1868 Hofmann B., Mathé P., 2018, *Inverse Problems*, **34**, 015007  
1869 Ilbert O., et al., 2006, *A&A*, **457**, 841  
1870 Ivezić Ž., et al., 2008, preprint (arXiv:0805.2366),  
1871 Izbicki R., Lee A. B., 2017, *Electron. J. Statist.*, **11**, 2800  
1872 Izbicki R., Lee A. B., Freeman P. E., 2017, *Ann. Appl. Stat.*,  
1873      11, 698  
1874 Laigle C., et al., 2016, *ApJS*, **224**, 24  
1875 Laureijs R., et al., 2011, preprint (1110.3193),  
1876 Leistedt B., Hogg D. W., 2017, *ApJ*, **838**, 5  
1877 Maddox N., Hewett P. C., Warren S. J., Croom S. M., 2008,  
1878      *MNRAS*, **386**, 1605  
1879 Malz A., Hogg D., in prep., CHIPPR, chippr  
1880 Malz A., Marshall P., DeRose J., Graham M., Schmidt S., Wechsler R., 2018, *AJ*, Accepted,  
1881  
1882 Mandelbaum R., et al., 2008, *MNRAS*, **386**, 781  
1883 Massarotti M., Iovino A., Buzzoni A., 2001, *A&A*, **368**, 74  
1884 Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes  
1885      J. D., Castander F. J., Paltani S., 2017, *ApJ*, **841**, 111  
1886 Newman J. A., 2008, *ApJ*, **684**, 88  
1887 Newman J. A., et al., 2015, *Astroparticle Physics*, **63**, 81  
1888 Polsterer K. L., D'Isanto A., Gieseke F., 2016, preprint  
1889      (arXiv:1608.08016),  
1890 Rasmussen C., Williams C., 2006, *Gaussian Processes for Ma-  
1891 chine Learning. Adaptative computation and machine learn-  
1892 ing series*, MIT Press, Cambridge, MA  
1893 Rau M. M., Seitz S., Brimiouille F., Frank E., Friedrich O.,  
1894 Gruen D., Hoyle B., 2015, *MNRAS*, **452**, 3710  
1895 Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S.,  
1896 2013, *ApJ*, **771**, 30  
1897 Riccio G., Brescia M., Cavaudi S., Mercurio A., di Giorgio A.,  
1898 Molinari S., 2017, *PASP*, 129  
1899 Richards G. T., et al., 2006, *ApJS*, **166**, 470  
1900 Richards G. T., et al., 2015, *ApJS*, **219**, 39  
1901 Sadeh I., Abdalla F. B., Lahav O., 2016, *PASP*, **128**, 104502  
1902 Salvato M., et al., 2011, *ApJ*, **742**, 61