

An assessment of photometric redshift PDF performance in the context of LSST

S.J. Schmidt¹, A.I. Malz^{2,3}, J.Y.H. Soo⁴, M. Brescia⁵, S. Cavaudi^{5,6}, G. Longo⁶, I.A. Almosallam^{7,8}, M.L. Graham⁹, A.J. Connolly⁹, E. Nourbakhsh¹, J. Cohen-Tanugi¹⁰, H. Tranin¹⁰, P.E. Freeman¹¹, K. Iyer¹², J.B. Kalmbach¹³, E. Kovacs¹⁴, A.B. Lee¹¹, C. Morrison⁹, J. Newman¹⁵, E. Nuss¹⁰, T. Pospisil¹¹, M.J. Jarvis^{16,17}, R. Izbicki^{18,19}

(LSST Dark Energy Science Collaboration)

¹ Department of Physics, University of California, One Shields Ave., Davis, CA, 95616, USA

² Center for Cosmology and Particle Physics, New York University, 726 Broadway, New York, 10003, USA

³ Department of Physics, New York University, 726 Broadway, New York, 10003, USA

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

⁵ INAF-Capodimonte Observatory, Salita Moiariello 16, I-80131, Napoli, Italy

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

⁷ King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

⁸ Information Engineering, Parks Road, Oxford, OX1 3PJ, UK

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

¹⁰ Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS, Montpellier, France

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

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

¹³ Department of Physics, University of Washington, Box 351560, Seattle, WA 98195, USA

¹⁴ Argonne National Laboratory, Lemont, IL 60439, USA

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

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

¹⁷ Department of Physics and Astronomy, University of the Western Cape, Bellville 7535, South Africa

¹⁸ Department of Statistics, Federal University of Sao Carlos, Sao Carlos, Brazil

¹⁹ External collaborator

12 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 including the Kolmogorov-Smirnov statistic, Cramer-von Mises statistic, Anderson-Darling statistic, Kullback-Leibler divergence, $N(z)$ moments, quantile-quantile plots and probability integral transform. 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), 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)¹ lists the photometric redshift goals for a magnitude limited sample with $i < 25$ as: root-mean-square error with a goal of $\sigma_z < 0.02(1+z)$; 3σ “catastrophic outlier” rate below 10%; bias below 0.003².

The tremendous size of LSST’s galaxy catalog 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. In addition to the contribution of low signal-to-noise photometry to the systematic error of photo- z s, these populations introduce major physical degeneracies, for example the Lyman break/Balmer break degeneracy, that were not present in the populations covered in previous 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). In order to meet LSST’s demanding error budget, it will be necessary to fully characterize those degeneracies wherein multiple redshift solutions have comparable likelihood.

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 (PDF), 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 rel-

ative χ^2 values of template spectra to likelihoods to estimate $p(z)$. Soon after, codes such as Benítez (2000) added a Bayesian prior and output a posterior probability distribution. While many early machine learning based algorithms focused on a point-estimate, Firth et al. (2003) used a neural net with 1000 realizations scattered within the photometric errors to estimate a $p(z)$. As more groups began to employ photometric redshifts in their cosmological analyses, 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).

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 photometric bands. DES compared several codes for point estimates (Sánchez et al. 2014) and a summary statistic of photo- z interim posteriors for tomographically binned galaxy subsamples (Bonnett et al. 2016). This paper is distinguished by its inclusion of 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 in this work are a major focus of the Photometric Redshift working group of the LSST Dark Energy Science Collaboration (LSST-DESC). This work is laid out in the Science Roadmap (SRM)³ as one of the critical activities to be completed in preparation for dark energy science analysis on the first year LSST data. This is the first of multiple papers by the working group, which will grow in sophistication. In this initial paper we focus on evaluating the performance of photometric redshift codes and their ability to produce accurate PDFs in the presence of representative training sets. This can be thought of as an initial test under near perfect conditions, before further complexities are added in future papers. 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.

Certain science cases need redshift information on individual objects, e. g. identification of host galaxy redshift for supernova classification, or identifying potential cluster membership. Other science cases need only ensemble redshift information; for instance current weak lensing techniques require the overall redshift distribution $N(z)$ for tomographic redshift samples, but do not need single galaxy estimates. In the case of the multiple types of probes of cosmology enabled by the LSST cosmology sample of several billion galaxies, the number of redshift bins and their photo-

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

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

³ Available at: http://lsst-desc.org/sites/default/files/DESC_SRMs_V1_1.pdf

113 z requirements vary with the specific probe; 2-point angular correlations benefit from many bins, while weak lensing 114 probes do not (due to the wide lensing kernel). Large photometric surveys such as LSST must develop algorithms that 115 meet the needs of all such science cases. In order to meet 116 these ambitious goals for photo- z accuracy, every aspect of 117 photo- z estimation will have to be optimized: the algorithms 118 employed, both template and machine-learning based (both 119 in design and implementation); the spectroscopic data used 120 as a training set for machine learning algorithms or to es- 121 timate template sets and train Bayesian priors; and proba- 122 bilistic catalog compression schemes that balance informa- 123 tion retention against limited storage resources. 124

125 Before moving forward, we must address how the best 126 methods may be unique to the performance metrics and sci- 127 ence cases considered and what distinguishes photo- z PDFs 128 of different methods from one another. Though photo- z 129 PDFs are often written simply as $p(z)$, the PDF itself must 130 be an interim posterior distribution $p(z|d, I)$, the probability 131 of redshift conditioned on photometric data d that has actu- 132 ally been observed and the prior information I that guides 133 how a redshift is extracted from the photometry. If we run 134 multiple photo- z codes on a single dataset, the photo- z in- 135 terim posteriors will not be identical because each code is 136 based on assumptions in the form of an interim prior — 137 these assumptions form the premise for photo- z estimation 138 as a whole and are the only way to introduce differences in 139 estimates of what would otherwise be a shared photo- z pos- 140 terior $p(z|d)$ regardless of the code used to obtain it. Though 141 explicit knowledge of the interim prior is necessary to use 142 photo- z interim posteriors self-consistently in physical infer- 143 ence, the interim prior of a particular methodology is often 144 implicit and not necessarily shared among all galaxies in the 145 catalog.

146 This paper therefore aims (1) to constrain the impact of 147 the interim prior I by separating it into a component I_H re- 148 presenting the method itself and a component I_D representing 149 physical information, such as a training set or SED template 150 library and (2) to present a procedure for evaluating the 151 performance of photo- z codes in generic tests that may include 152 many more systematics in the interim prior I . In order to 153 isolate the effects encapsulated by I_H of variation between 154 codes from issues with the training set or template library 155 encapsulated by I_D , we use an identical set of simulated 156 galaxies for every code and construct a template library 157 and training sample that are *complete and representative* 158 and shared among all codes; that is, our training sample for 159 machine learning codes is drawn from the same underlying 160 galaxy population as our test set, with no additional selec- 161 tions, and the SED library used for template-based codes is 162 the same as the one used to generate the photometric data. 163 We explore a number of performance metrics in this paper, 164 not to make a conclusion regarding the superiority or even 165 relative favorability of each code but to establish a method 166 for comparing photo- z PDFs derived by different methods. 167 These test conditions set the stage for addressing in a future 168 paper the crucial issue of incomplete and non-representative 169 prior information. 170

171 The outline of the paper is as follows: in § 2 we present 172 the simulated data set; in § 3 we describe the current genera- 173 tion codes employed in the paper; in § 4 we discuss the inter- 174 pretation of photo- z PDFs in terms of metrics of accuracy;

175 in § 5 we show our results and compare the performance of 176 the codes; in § 6 we offer our conclusions and discuss future 177 extensions of this work.

178 2 THE SIMULATION AND MOCK GALAXY CATALOG

180 In order to test the current generation codes, we employ 181 a simulated galaxy catalogue. The simulation is completely 182 catalogue-based, with no image construction or mock mea- 183 surements made. We describe these in detail below.

184 2.1 Buzzard-v1.0 simulation

185 The BUZZARD-HIGHRES-V1.0 186 put in cites to in prep 187 Buzzard 188 papers catalogue construction started with a dark matter 189 only simulation. This N-body simulation contained 2048^3 190 particles in a 400 Mpc h^{-1} box. 191 [N] snapshots (with smooth- 192 ing and interpolation between snapshots) were saved in 193 order to construct a lightcone. Dark matter halos were identi- 194 fied using the ROCKSTAR software package (195 Behroozi et al. 196 2013). These dark matter halos were populated with galax- 197 ies with a stellar mass and absolute r -band magnitude in 198 the SDSS system determined using a sub-halo abundance 199 matching model constrained to match both projected two- 200 point galaxy clustering statistics and an observed condi- 201 tional stellar mass function (202 Reddick et al. 203 2013).

202 To assign an SED to each galaxy, the *Adding Den- 203 sity Dependent Spectral Energy Distributions* (ADDSEDS, 204 deRose in prep.)⁴ procedure was used. This consisted of 205 training an empirical relation between absolute r -band mag- 206 nitude, local galaxy density, and SED using a sample of 207 $\sim 5e^5$ galaxies from the magnitude-limited Sloan Digital 208 Sky Survey Data Release 6 Value Added Galaxy Catalog 209 (210 Blanton et al. 211 2005)[Note: is this the proper reference to 212 SDSS-NYU VAGC? File is called combined_dr6.cooper.fits, 213 but I don't see which Cooper et al 2006 this is supposed 214 to refer to?]. Each SDSS spectrum is fit with a sum of five 215 SED components using the K-CORRECT v?216 software pack- 217 age⁵ (218 Blanton & Roweis 219 2007), thus each galaxy SED is 220 parameterized as five weights for the basis SEDs. The distance 221 to the spatial projected fifth-nearest neighbour was used as a 222 proxy for local density in the SDSS training sample. For each 223 simulated galaxy, a “random” 224 [details] galaxy with “similar” 225 [details] absolute r -band magnitude and local galaxy density 226 was chosen from the training set, and that training galaxy’s 227 SED was assigned to the simulated galaxy. Given the SED, 228 absolute r -band magnitude and redshift, we computed ap- 229 parent magnitudes in the six LSST filter passbands, $ugrizy$. 230 We assigned magnitude errors in the six bands using the 231 simple model described in 232 Ivezić et al. (2008), assuming full 233 10-year depth observations had been completed. The num- 234 ber of total 30-second visits assumed when generating the 235 photometric errors differs slightly from the fiducial numbers 236 assumed for LSST: we assume 60 visits in u-band, 80 vis- 237 its in g-band, 180 visits in r-band, 180 visits in i-band, 160 238 visits in z-band, and 160 visits in y-band.

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

240 ⁵ <http://kcorrect.org>

4 LSST Dark Energy Science Collaboration

228 2.1.1 Selection of training and test sets

229 The total catalogue covered 400 square degrees and con-
 230 tained 238 million galaxies to an apparent magnitude limit
 231 of $r = 29$ and spanning the redshift range $0 < z \leq 8.7$. This
 232 catalogue contained two orders of magnitude more galaxies
 233 than were needed for this study, so only ~ 8 square degrees
 234 were used. Systematic problems with galaxy colors above
 235 $z > 2$ were observed, so the catalogue was trimmed to in-
 236 clude only galaxies in the redshift range $0 < z \leq 2.0$. A
 237 random subset of the the remaining galaxies was chosen,
 238 and placed at random into either a “training” set (10 per
 239 cent of the sample), for which the galaxies true redshifts
 240 will be supplied, or a “test” set (the remaining 90 per cent
 241 of the sample), for which each code will need to predict a
 242 redshift PDF for each galaxy. The resulting catalogues con-
 243 tain 111 171 training galaxies and 1 000 883 test galaxies. We
 244 restrict our analysis to a sample with $i < 25.3$, which give a
 245 signal-to-noise ~ 30 for most galaxies, a cut often referred
 246 to as the expected “LSST Gold Sample”. This magnitude
 247 cut results in a training set with 44 404 galaxies and a test
 248 set containing 399 356 galaxies. All subsequent results will
 249 evaluate this “gold sample” test set.

250 2.1.2 Templates

251 As mentioned in Section 2.1, the SEDs in the Buzzard sim-
 252 ulation are drawn from an empirical set of SEDs taken from
 253 the SDSS DR6 NYU-VAGC, a sample of roughly $\sim 5e^5$
 254 galaxies with spectra in SDSS. To determine a finite set of
 255 templates to use with template fitting codes we take the five
 256 SED weight coefficients for each of the $\sim 500\,000$ galaxies in
 257 the SDSS sample and run a simple K-means clustering algo-
 258 rithm on this five dimensional space. The K-means cluster
 259 centres span the space of coefficients and properly reflect the
 260 underlying density in the coefficient space, thus providing a
 261 reasonable approximation for a spanning SED set. An ad-
 262 hoc number of $K = 100$ was chosen and the 100 K-means
 263 centre positions are taken as the weights for the k-CORRECT
 264 SED components to construct one hundred template SEDs.
 265 These 100 templates were provided, however not every tem-
 266 plate code uses this set of one hundred templates: because
 267 EAZY was designed and written to use the same five basis
 268 templates employed by k-CORRECT when constructing our
 269 mock galaxies, EAZY was run using linear combinations of
 270 these five templates rather than using the 100 discrete tem-
 271 plates.

272 2.1.3 Limitations

273 For our initial investigation of photometric redshift codes,
 274 we begin with a data set that is somewhat idealized, and
 275 does not contain all of the complicating factors present in
 276 real data. In several cases, the simplification is done with
 277 a purpose, with potentially confounding effects excluded
 278 in order to better isolate the differences between current-
 279 generation photo- z codes, and their causes. We list several
 280 of the simulations limitations in this section. As the sim-
 281 ulation is catalogue-based, no image level effects, such as
 282 photometric measurement effects, object blending, contami-
 283 nation from sky background (Zodiacal light, scattered light,

284 etc...), lensing magnification, or Galactic reddening are in-
 285 cluded. No stars are included in the catalogue, nor are the
 286 effects of AGN. As all SEDs are constructed from only five
 287 basis templates, properties of the galaxy population will be
 288 restricted to follow linear combinations of the characteristics
 289 of the five basis templates, so certain non-linear features, for
 290 example the full range of emission line fluxes relative to the
 291 continuum, will not be included in the model galaxy pop-
 292 ulation. No additional dust reddening intrinsic to the host
 293 galaxy is included, the only approximation of dust extinc-
 294 tion comes in the form of dust encoded in the five basis SEDs
 295 via the training set used to create the basis templates. Sim-
 296 ple linear combinations of these basis templats will, once
 297 again, not explore the full range of realistic dust extinction
 298 observed in galaxy populations.

299 3 METHODS

300 Here we outline the photo- z PDF codes tested in this study.
 301 In total, eleven distinct codes are tested. This sample is not
 302 comprehensive, codes were chosen based on the expertise
 303 available within the group; however, those chosen do cover
 304 a broad range of the current-generation methods used in the
 305 field. Both template-based and machine learning approaches
 306 are included and each are described separately in Secs. 3.1
 307 and 3.2 respectively. The list of codes are summarized in Ta-
 308 ble. 1. All code runners were asked to output redshift pos-
 309 terior estimates on 200 linear-spaced bins between redshifts
 310 0 and 2.

311 The questions that must be answered for each code are:
 312 what unique features are included in the specific implemen-
 313 tation that influence the output $p(z)$. What form of valida-
 314 tion was performed with the training data, how were pho-
 315 tomometric uncertainties employed in the analysis, how were
 316 negative fluxes treated, what specific prior form was em-
 317 ployed (for template based codes), or what specific machine
 318 learning architecture was used (for ML codes)?

319 3.1 Template-based Approaches

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

$$327 \chi^2(z, T, A) = \sum_i^{N_{\text{filt}}} \left(\frac{F_{\text{obs}}^i A \times F_{\text{pred}}^i(T, z)}{\sigma_{\text{obs}}^i} \right)^2 \quad (1)$$

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

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

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

3.1.1 BPZ

BPZ⁶ (Bayesian Photometric Redshift, Benítez 2000) is a template-based photo- z code that compares the expected colors (C) calculated for a set of spectral energy distribution (SED) types/templates (T) to the observed colors to calculate the likelihood of observing colors at each redshift for each type, $p(C|z, T)$. The likelihoods at each redshift are related to the χ^2 in Equation 1 by the simple form: likelihood $\propto e^{-\chi^2/2}$. The code employs an empirically determined Bayesian prior in apparent magnitude (m_0) and SED-type. Assuming that the SED-types are spanning and exclusive, we can determine the redshift posterior $p(z|C, m_0)$ by marginalizing over all SED-types with a simple sum (Eq. 3 from Benítez 2000):

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

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

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

generated by setting the parameter PROBS_LITE=TRUE in the BPZ parameter file.

3.1.2 EAZY

EAZY⁷ (Easy and Accurate Photometric Redshifts from Yale, Brammer et al. 2008) is a template-based photo- z code that includes several features that extend the basic χ^2 fit used in many template codes. The code can fit the observed photometry with SEDs created from a linear combination of a set of templates at each redshift, and the best-fit SED is found by simultaneously fitting one, two or all of the templates by minimizing χ^2 . The minimized $\chi^2(z)$ is then combined with an apparent magnitude prior to obtain the posterior redshift probability distribution. On examination of the source code for EAZY, it appears that rather than marginalizing across all templates in the χ^2 calculation, EAZY takes only the minimum value of χ^2 at each redshift. This improper marginalization does not lead to the correct posterior distribution, an implementation issue that will need to be addressed in the future. EAZY can also account for the uncertainties in the templates by adding an empirically derived template error in quadrature as a function of redshift to the flux errors.

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

3.1.3 LePhare

LEPHARE⁸ (Photometric Analysis for Redshift Estimate, Arnouts et al. 1999; Ilbert et al. 2006) is a photo- z reconstruction code based on a χ^2 template-fitting procedure. The observed colors are matched with the colours predicted from

⁶ <http://www.stsci.edu/~dcoe/BPZ/>

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

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

6 LSST Dark Energy Science Collaboration

406 a set of spectral energy distribution (SED) which can be ei-
 407 ther synthetic or based on a semi-empirical approach.

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

413 In this paper we use LEPHARE v 2.2. The set of tem-
 414 plates used for fitting the photo- z 's are the 100 discrete Buzz-
 415 zard SED templates as described in section 2.1.2, and the
 416 full $p(z)$ corresponds to the likelihoods calculated at each
 417 point on our z -grid.

418 3.2 Training-based Codes

419 The training-based codes use a variety of algorithms in or-
 420 der to estimate $p(z)$, specifics of each implementation are
 421 described in the subsections. Some aspects of data treat-
 422 ment were left to the individual code runners, for example,
 423 whether/how to split the available data with known redshifts
 424 into separate training and validation sets. Another key dif-
 425 ference is the treatment of non-detections in one or more
 426 bands. Some codes choose to ignore a band, others replace
 427 the value with either an estimate for the detection limit,
 428 the mean of other values in the training set, or another de-
 429 fault value. There are varying conventions among training-
 430 based codes for treatment of non-detections, and no one
 431 prescription dominates in the photo- z literature. The spe-
 432 cific choices for each code affect the results, and contribute
 433 to the implicit prior influencing their output. However, we
 434 remind the reader that only 2.0 per cent of our sample has
 435 non-detections, almost exclusively in the u-band, and thus
 436 should not dominate the code performance differences.

437 3.2.1 ANNz2

438 ANNz2⁹ (Sadeh et al. 2016) is a software package that has
 439 the ability to employ several machine learning algorithms,
 440 including artificial neural networks (ANN), boosted deci-
 441 sion tree (BDT) and k-nearest neighbour (KNN). Using
 442 the Toolkit for Multivariate Data Analysis (TMVA) with
 443 ROOT¹⁰, it can either run a single machine learning algo-
 444 rithm and return results, or it can run multiple algorithms
 445 simultaneously and output photo- z 's as a weighted combi-
 446 nation of the different algorithms. In this study, only ANNs
 447 were employed. The redshift PDFs are produced by run-
 448 ning an ensemble set of ANNs, each with different random
 449 seeds used in initialization of input parameters for train-
 450 ing. Uncertainties for each method are estimated from a
 451 KNN-uncertainty estimator (Oyaizu et al. 2008). The final
 452 PDF can either be the “best” of the candidate PDFs, or a
 453 weighted average of the PDFs based on their error estimates
 454 for each of the ensemble members.

455 In this study, ANNz2 v. 2.0.4 was used. A set of 5
 456 ANNs with architecture 6 : 12 : 12 : 1 (6 *ugriz* inputs, 2
 457 hidden layers with 12 nodes each, and 1 output) with differ-
 458 ent random seeds are used during each training. Half of the

459 training set is used as a validation set to prevent overtrain-
 460 ing. All training objects are set to have detected magnitudes,
 461 however the non-detections ($\text{mag} = 99$) in the testing set are
 462 replaced with the mean of that particular band.

463 3.2.2 Colour-Matched Nearest-Neighbours

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

$$473 D_M = \sum^{N_{\text{colours}}} \frac{(c_{\text{train}} - c_{\text{test}})^2}{(\delta c_{\text{test}})^2} \quad (3)$$

474 where c colour, δc_{test} is the measurement error on the colour,
 475 and N_{colours} is the total number of colors (i.e., in our case *u*-
 476 *g*, *g-r*, *r-i*, *i-z*, and *z-y*). Then, we choose a threshold value
 477 for D_M that defines the colour match set based on a set
 478 value of the percent point function (PPF): for example, for
 479 $N_{\text{dof}} = 5$, choosing a vPPF = 95 per cent of all training
 480 galaxies consistent with the test galaxy will have $D_M <$
 481 11.07 (where N_{dof} is the number of degrees of freedom, in
 482 this case the number of colours). If a galaxy had a non-
 483 detection in a band, that band was dropped and N_{dof} was
 484 reduced by one in the colour-matching space. For a given test
 485 galaxy, the $p(z)$ is the normalized distribution of the true
 486 catalogue redshifts of this colour-matched subset of training
 487 galaxies.

488 We have applied the nearest-neighbours colour-
 489 matching photometric redshift estimator described in G18
 490 to the simulated data. Compared to its application in G18,
 491 there are some minor differences in the application of this
 492 estimator to the Buzzard catalogue. First, we do not impose
 493 non-detections on galaxies with a magnitude fainter than
 494 the expected LSST 10-year limiting magnitude or bright
 495 enough to saturate with LSST: *all* of the photometry for
 496 all the galaxies in the test and training sets are used for this
 497 experiment. Second, as in G18 we do apply an initial cut
 498 in colour to the training set before calculating the Ma-
 499 halanobis distance in order to accelerate processing, and also
 500 use a magnitude pseudo-prior to improve photo- z estimates,
 501 but for both we have used different cut-off values that are
 502 appropriate for the Buzzard galaxies' colours and magni-
 503 tudes. Third, we set different parameters for the identifica-
 504 tion of the colour-matched subset of training galaxies and
 505 the selection of a photometric redshift estimate. In G18 we
 506 used a percent point function (PPF) value of 0.68 to identify
 507 the colour-matched subset of training galaxies and used the
 508 redshift of nearest neighbour in colour-space as the photo- z
 509 estimate. These choices work well when the desire is to ob-
 510 tain accurate photo- z estimates for most test-set galaxies,
 511 but does not return an accurate $p(z)$ in all cases – espe-
 512 cially for galaxies that are bright and/or have few matches
 513 in colour-space. Since an accurate estimate of $p(z)$ is desired
 514 for this work we make several changes to our implemen-
 515 tation of the CMNN photo- z estimator. We continue to use

⁹ <https://github.com/IftachSadeh/ANNZ>

¹⁰ <http://tmva.sourceforge.net/>

516 a percent point function of PPF = 0.95 to generate the
 517 subset of colour-matched training galaxies, but weight them
 518 by the inverse of their Mahalanobis distance. This weighting
 519 maintains some of the accuracy that was previously achieved
 520 by simply using the nearest neighbour in colour-space. We
 521 then use the weights to create the $p(z)$ instead of having the
 522 redshift of each colour-matched training-set galaxy count
 523 equally. To obtain a robust estimate of the $p(z)$ for galaxies
 524 with a small number of colour-matched training set galaxies,
 525 when this number is less than 20 the nearest 20 neighbours
 526 in colour-space are used instead, and we convolve the $p(z)$
 527 with a Gaussian with a standard deviation of:

$$528 \sigma = \sigma_{\text{train}} \sqrt{(\text{PPF}_{20}/0.95)^2 - 1} \quad (4)$$

529 to appropriately broaden it so that the $p(z)$ for these test
 530 galaxies represents the enlarged PPF value associated with
 531 it. Overall, these three changes will yield less precise photo- z
 532 estimates compared to those presented in G18, but they will
 533 all have significantly more accurate estimates of the $p(z)$,
 534 particularly for the brightest test galaxies. This is sufficient
 535 for this work because, as described in G18, the goal of the
 536 CMNN photo- z estimator was never to provide the “best”
 537 (or even competitive) estimates in the first place, given its
 538 reliance on a deep training set, but rather to provide a means
 539 for direct comparisons between LSST photometric quality
 540 and photo- z estimates. With this work we show how the
 541 input parameters should be set in order to return accurate
 542 $p(z)$ estimates in addition to point value estimates.

543 3.2.3 Delight

544 DELIGHT¹¹ (Leistedt & Hogg 2017) infers photo- z ’s by using
 545 a data-driven model of latent SEDs and a physical model of
 546 photometric fluxes as a function of redshift. Delight models
 547 the underlying latent SEDs as a linear combination of
 548 a set of pre-defined template SEDs, plus zero mean Gaus-
 549 sian processes with factorized kernels. Generally, machine
 550 learning methods rely on representative training data with
 551 similar band passes, while template based methods rely on
 552 a complete library of templates based on physical models
 553 constructed. DELIGHT is constructed in attempt to com-
 554 bine the advantages and eliminate the disadvantages of both
 555 template-based and machine learning algorithms: it con-
 556 structs a large collection of latent SED templates (or phys-
 557 ical flux-redshift models) from training data, with a template
 558 SED library as a guide to the learning of the model. The ad-
 559 vantage of DELIGHT is that it neither needs representative
 560 training data in the same photometric bands, nor does it
 561 need detailed galaxy SED models to work.

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

$$566 p(z|\hat{\mathbf{F}}) \approx \sum_i p(\hat{\mathbf{F}}|z, t_i) p(z|t_i)p(t_i), \quad (5)$$

567 where $p(z|t_i)p(t_i)$ captures prior information about the red-
 568 shift distributions and abundances of the galaxies, with t_i

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

569 denoting the galaxy template; while $p(\hat{\mathbf{F}}|z, t_i)$ is the poste-
 570 rior of noisy flux $\hat{\mathbf{F}}$ at redshift z . For each training-target
 571 pair, $p(\hat{\mathbf{F}}|z, t_i)$ is evaluated as follows:

$$572 p(\hat{\mathbf{F}}|z, t_i) = \int p(\hat{\mathbf{F}}|\mathbf{F}) p(\mathbf{F}|z, z_i, \hat{\mathbf{F}}_i) d\mathbf{F}, \quad (6)$$

573 where $p(\hat{\mathbf{F}}|\mathbf{F})$ is the likelihood function, it compares the
 574 noisy real flux $\hat{\mathbf{F}}$ with the noiseless flux \mathbf{F} obtained from the
 575 linear combination of template models, carefully constructed
 576 to account for model uncertainties and different normaliza-
 577 tion of the same SED; while $p(\mathbf{F}|z, z_i, \hat{\mathbf{F}}_i)$ is the prediction
 578 of flux at a different redshift z with respect to the training
 579 object with redshift z_i and flux $\hat{\mathbf{F}}_i$. Eq. 6 is essentially the
 580 probability that the training and the target galaxies having
 581 the same SED but at a different redshift. The flux prediction
 582 $p(\mathbf{F}|z, z_i, \hat{\mathbf{F}}_i)$ of the training galaxy at redshift z is modeled
 583 via a Gaussian process,

$$584 F_b \sim \mathcal{GP} \left(\mu^F, k^F \right), \quad (7)$$

585 with mean function μ^F and kernel k^F , both imposed to
 586 capture expected correlations resulting from the known un-
 587 derlying physics (i.e., fluxes resulting from observing SEDs
 588 through filter response, and the SEDs being redshifted). The
 589 reader should refer to Leistedt & Hogg (2017) for further de-
 590 tails.

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

600 3.2.4 FlexZBoost

601 FLEXZBOOST¹² (Izbicki & Lee 2017) is a particular realiza-
 602 tion of FlexCode, which is a general-purpose methodology
 603 for converting any conditional mean point estimator of z to
 604 a conditional density estimator $f(z|\mathbf{x})$, where \mathbf{x} here repre-
 605 sents our photometric covariates and errors.¹³ The key idea
 606 is to expand the unknown function $f(z|\mathbf{x})$ in an orthonormal
 607 basis $\{\phi_i(z)\}_i$:

$$608 f(z|\mathbf{x}) = \sum_i \beta_i(\mathbf{x}) \phi_i(z). \quad (8)$$

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

$$611 \beta_i(\mathbf{x}) = \mathbb{E} [\phi_i(z)|\mathbf{x}] \equiv \int f(z|\mathbf{x}) \phi_i(z) dz. \quad (9)$$

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

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

8 LSST Dark Energy Science Collaboration

These coefficients can easily be estimated from data by regression.

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

3.2.5 GPz

GPz¹⁴ (Almosallam et al. 2016a,b) is a sparse Gaussian process based code, a scalable approximation of full Gaussian Processes (Rasmussen & Williams 2006), with the added feature of being able to produce input-dependent variance estimations (heteroscedastic noise). The model assumes that the probability of the output y , the redshift, given the input x , the photometry, is $p(y|x) = \mathcal{N}(y|\mu(x), \sigma(x)^2)$. The mean function, $\mu(x)$, and the variance function $\sigma(x)^2$ are both linear combinations of basis functions that take the following form:

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

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

A unique feature in GPz, is that the variance estimate is composed of two terms each quantifying a different source of uncertainty. One term (the model uncertainty) reflects how much of the uncertainty is due to lack of training samples at the location of interest, whereas the second term (the noise uncertainty) reflects how much of the uncertainty is caused from observing many noisy samples at that location. Thus, the predictive variance can determine whether we need more representative samples or more precise samples for any particular location in the input space. GPz can also emphasize the importance of some samples as weights. This weight can be for example $|z_{\text{spec}} - z_{\text{phot}}|/(1 + z_{\text{spec}})$ 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 any missing magnitude values using a simple linear model that estimates the missing magnitudes given the observed ones. The log transformation helps to smooth the long tail distribution of the magnitude errors, which is more stable numerically and makes the optimization process unconstrained. The missing values are imputed by computing the mean of the training set μ and its covariance Σ , then we use the following equation to estimate the missing values from the observed ones

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

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

3.2.6 METAPhOR

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

METAPhOR includes data modules for pre-processing, photo-z estimation, and PDF estimation, and post-processing. The pre-processing includes a model for perturbation of the photometry that is employed in calculating the PDF of the photo-z estimation errors. The photometric perturbation is defined as: $m_{ij} = m_{ij} + \alpha_i F_{ij} * u_{\mu=0, \sigma=1}$, where α_i is a user selected multiplicative constant (useful in case of multi-survey photometry), $u_{\mu=0, \sigma=1}$ is a random value from the standard normal distribution and F_{ij} is a bimodal function (a constant function + polynomial fitting of the mean magnitude errors on the binned bands), heuristically tuned in such a way that the constant component is the threshold under which the polynomial function is considered too low to provide a significant noise contribution to the photometry perturbation.

As main prerogative, METAPhOR is able to provide a PDF for ML methods by taking into account the photometric errors provided with data, by running N trainings on the same training set, or M trainings on M different random extractions from the KB. The different test sets, used to produce the PDF, are thus obtained by introducing a proper perturbation, parametrized from the photometric error distribution in each band, on the photometric data populating the original test set (Brescia et al. 2018). For the

¹⁴ <https://github.com/OxfordML/GPz>

present work since it was required to produce a redshift (and a PDF) for each object of the test set we decided to apply a hierarchical kNN to replace the missing detections with values based on their neighbors. The reliability of PDFs and point estimation is lower. No cross validation has been used.

3.2.7 SkyNet

SKYNET¹⁵ (Graff et al. 2014) is a publicly available neural network software, based on a 2nd order conjugate gradient optimization scheme (see Graff et al. 2014, for further details).

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

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

3.2.8 TPZ

TPZ¹⁶ (Trees for Photo-z, Carrasco Kind & Brunner 2013; Carrasco Kind & Brunner 2014) is a parallel machine learning algorithm that generates photometric redshift PDFs using prediction trees and random forest techniques. The code recursively splits the input data (i. e. the training sample), into two branches, one after another, until a terminal leaf is created that meets a termination criterion (e. g. a minimum leaf size or a variance threshold). Bootstrap samples from the training data and associated errors are used to build a set of prediction trees. In order to minimize correlation between the trees, the data is divided in such a way that the highest information gain among the random subsample of features is obtained at every point. The regions in each terminal leaf node corresponds to a specific subsample of the entire data that possesses similar properties.

The training data is examined before running TPZ. Since TPZ does not handle non-detections (magnitudes flagged as 99.0), we replace these values with an approximation of the 1σ detection threshold, i. e. a signal to noise ratio of 1 in terms of magnitude uncertainty using the equation $dm = 2.5 \log(1 + N/S)$ where $dm \sim 0.7526$ mag

for $N/S = 1$. That is, for each band, we replace the non-detection with the magnitude corresponding to the error of 0.7526 from the error model forecasted for 10-year LSST data. The Out-of-Bag (Breiman et al. 1984; Carrasco Kind & Brunner 2013) cross-validation technique is used within TPZ to evaluate its predictive validity and determine the relative importance of the different input attributes. We employed this information to calibrate our algorithm.

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

3.3 Simple Ensemble Estimator

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

$$p(z) = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} z_{\text{train}} \quad (12)$$

That is, we simply set the redshift PDF of every galaxy equal to the normalized $N(z)$ of the training sample. This estimator is essentially a k nearest-neighbour estimator with k equal to the number of galaxies in the training sample. As the training sample is drawn from the same underlying distribution as the test sample, modulo small deviations due to sample size, the quantiles of the training and test distributions should be identical, modulo fluctuations due to finite sample size. This is a wildly unrealistic estimator, as it assigns all galaxies, no matter their apparent magnitude, colour, or true redshift, the same redshift PDF, and is thus uninformative at the level of individual object redshifts, but is designed to perform very well for the ensemble of all objects. If the training set was not representative, this estimator would produce biased results, and any attempts to break up the sample into tomographic bins will fail, as every galaxy has an identical $p(z)$. We will discuss this method and cautions relative to metrics in Section 5.3.

4 METRICS FOR QUANTIFYING PDF COMPARISONS

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

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

¹⁶ <https://github.com/mgckind/MLZ>

$$\hat{p}_j(z_i) = p(z|d_i, I_D, I_H), \quad (13)$$

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

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

$$\frac{\hat{p}_{i,H}(z)}{\hat{p}_{i,H'}(z)} \approx \frac{p(z|d_i, I_H)}{p(z|d_i, I_{H'})}, \quad (14)$$

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

As mentioned previously, there are cosmology probes that require knowledge of individual galaxy $p(z)$ and others that require only knowledge of the ensemble redshift distribution, $N(z)$. Due to the paucity of principled techniques for using and validating photo-z PDFs, there are few alternatives to the common practice of reducing photo-z PDFs to point estimates. Though this practice should not be encouraged, we also calculate traditional metrics based on the most common point estimators derived from photo-z PDFs. Those seeking to establish a connection to traditional ways of thinking about redshift estimation may consult the Appendix for these results.

There are a number of metrics that can be used to test the accuracy of a photo-z interim posterior as an estimator of a true photo-z posterior if it is known. Even for simulated data, the true photo-z PDF is in general not accessible unless the redshifts are in fact drawn from the true photo-z PDFs, a mock catalogue generation procedure that has not yet appeared in the literature. Furthermore, only limited applications of photo-z PDFs that could be used as the basis for a metric have been presented in the literature. The most popular application by far is the calculation of the overall redshift distribution $N(z)$, the true value of which is known for the BUZZARD simulation and will be denoted as $N'(z)$. Though alternatives exist (Malz & Hogg prep), stacking according to

$$\hat{N}^H(z) \approx \frac{1}{N_{tot}} \sum_i^{N_{tot}} \hat{p}_i^H(z) \quad (15)$$

is the most widely accepted method for estimating the redshift distribution from photo-z PDFs. If we assume that the response of estimators of $N(z)$ is uniform across all approaches H , then we may interpret metrics on the accuracy of $\hat{N}(z)$ obtained in this way. We must note, however, that this is a poor assumption in general. Under the setup of this

paper, the true redshift distribution $N'(z) = p(z|I_D)$ (i.e. because our training data is representative, the interim prior is the truth). In this ideal case, the method that would give the best approximation to $N'(z)$ would be one that neglects all the information contained in the photometry $\{d_i\}_{N_{tot}}$ and gives every galaxy the same photo-z PDF $\hat{p}_i(z) = N'(z)$ for all i . This is the exact estimator, TRAINZ, that we have described in Section 3.3, and which will serve as a point of reference for the other codes.

The exact implementation of the stacked estimator $\hat{N}^H(z)$ will depend on the parametrization of the photo-z PDFs, which may differ across codes and can affect the precision of the estimator (Malz et al. 2018); even considering a single method under the same parametrization, say a piecewise constant function over bins or a set of samples from the posterior, an estimator using $2N$ bins or samples will trivially be more precise than an estimator using N bins or samples. In order to minimize the effects of such choices, we asked those running all eleven codes to output $p(z)$ parameterized with a generous ≈ 200 piecewise constant bins spanning $0 < z < 2$. The piecewise constant format is chosen because of its established presence in the literature, and the choice of 200 bins was motivated by the approximate number of columns expected to be available for storage of $p(z)$ for the final LSST Project tables.¹⁷ All $p(z)$ catalogues are processed using the QP software package (Malz et al. 2018)¹⁸ for manipulating and calculating metrics of 1-dimensional PDFs. We will discuss the choice of $p(z)$ parameterization further in Section 5.

4.1 Metrics of an ensemble of photo-z interim posteriors

4.1.1 Probability integral transform (PIT)

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

$$\text{PIT} = \int_{-\infty}^{z_{\text{true}}} p(z) dz. \quad (16)$$

The distribution of PIT values quantifies the behavior of the ensemble of photo-z PDFs, enabling us to evaluate whether the $p(z)$ is, on average, accurate: The PIT value is the Cumulative Distribution Function (CDF) of the $p(z)$ evaluated at the true redshift. A catalogue of photo-z PDFs that are accurate should have a flat PIT histogram (i.e., the individual PIT values as samples from each CDF should match a Uniform(0,1) distribution if the CDFs are accurate). Specific deviations from flatness indicate inaccuracy: overly broad photo-z PDFs would manifest as underrepresentation of the lowest and highest PIT values, whereas overly narrow photo-z PDFs would manifest as over-representation of the lowest and highest PIT values. High frequency at only $\text{PIT} \approx 0$ and $\text{PIT} \approx 1$ indicates the presence of catastrophic outliers with highly inaccurate photo-z PDFs where the true redshift is outside of the support of $p(z)$. Tanaka et al. (2017) use the histogram of PIT values as a diagnostic indicator of

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

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

overall code performance, while Freeman et al. (2017) independently define the PIT and demonstrate how its individual values may be used both to perform hypothesis testing (via, e.g., the KS, CvM, and AD tests; see below) and to construct quantile-quantile plots.

4.1.2 Quantile-quantile (QQ) plot

The quantile-quantile (QQ) plot is a graphical method for comparing two distributions, where the quantiles of one distribution are plotted against the quantiles of the other distribution (A quantile being defined by partitioning a distribution into consecutive intervals containing equal amounts of probability, or equal numbers of objects in each interval). In this paper we show the quantiles of the PIT values compared 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. 15) 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:

$$L(f, \hat{f}) = \int \int (f(z|\mathbf{x}) - \hat{f}(z|\mathbf{x}))^2 dz dP(\mathbf{x}) \quad (17)$$

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

$$\mathbb{E}_{\mathbf{x}} \left[\int \hat{f}(z|\mathbf{X})^2 dz \right] - 2\mathbb{E}_{\mathbf{x}, z} [\hat{f}(Z|\mathbf{X})] + K_f, \quad (18)$$

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. 15.

4.2.1 Root-mean-square error (RMSE)

We employ the familiar root-mean-square error:

$$\text{RMSE} = \sqrt{\int_{-\infty}^{\infty} (\hat{f}(z) - f'(z))^2 dz}, \quad (19)$$

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_{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:

$$N_{\text{KS}} = \max_z (|F_{\text{phot}}(z) - F_{\text{spec}}(z)|). \quad (20)$$

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

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

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

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

$$A^2 = n \int_{-\infty}^{+\infty} \frac{(F_{\text{meas.}}(x) - F_{\text{ideal}}(x))^2}{F_{\text{ideal}}(x)(1 - F_{\text{ideal}}(x))} dF_{\text{ideal}} \quad (22)$$

where n is the sample size.

1029 **4.2.3 Moments**

1030 For the $\hat{N}(z)$ distributions we additionally calculate the first
 1031 three moments of the estimated redshift distribution for each
 1032 code and compare them to the moments of the true redshift
 1033 distribution $N'(z)$. The m th moment of a distribution is
 1034 defined as

$$1035 \langle z^m \rangle = \int_{-\infty}^{\infty} z^m N(z) dz. \quad (23)$$

1036 Here, we use the moments of the stacked estimator of the
 1037 redshift distribution function as the basis for a metric. The
 1038 closer the moments of $\hat{N}(z)$ for a photo-z PDF method are
 1039 to the moments of the true redshift distribution function
 1040 $N'(z)$, the better the photo-z PDF method.

1041 **5 RESULTS**1042 **5.1 Ensembles of photo-z interim posteriors**

1043 Fig. 1 Shows the $p(z)$ produced by each of our eleven photo-
 1044 z codes for four example galaxies which exemplify some
 1045 prominent cases that arise when estimating photo-z PDFs: a
 1046 narrow, unimodal redshift solution, a broader unimodal so-
 1047 lution, a bimodal distribution, and a complex, multimodal
 1048 distribution. The red vertical line represents the true red-
 1049 shift of the individual galaxy, and the blue curve repre-
 1050 sents the redshift probability. Several features are obvious
 1051 even in these illustrative examples. ANNz2, METAPHOR,
 1052 NN, and SKYNET all show an excess of small-scale features,
 1053 which appear to be print-through of the underlying train-
 1054 ing set galaxies. GPZ (in its current implementation), on
 1055 the other hand, always produces a single Gaussian, which
 1056 broadens to cover the multi-modal redshift solutions seen in
 1057 other codes.

1058 As stated in Section 4, $p(z)$ is parameterized as ≈ 200
 1059 piecewise constant bins covering $0 < z < 2$ for all eleven
 1060 codes, giving a grid size of roughly $\delta z = 0.01$ for each code.
 1061 A piecewise constant grid was a natural choice for some
 1062 photo-z codes, for instance most template-based codes com-
 1063 pute likelihoods on a fixed grid. In contrast, FlexZBoost, for
 1064 example, can return estimates on any grid without compres-
 1065 sion errors as its a basis expansion method where only the
 1066 expansion coefficients need to be stored. Codes with a na-
 1067 tive output format other than the shared piecewise constant
 1068 binning scheme (or one that can be losslessly converted to
 1069 it) may suffer from loss of information when converting to
 1070 it, which could artificially favor some codes over others.

1071 Furthermore, the fidelity of photo-z interim posteriors
 1072 in this format varies with the quality of the photometry. For
 1073 faint galaxies, this redshift resolution is sufficient to capture
 1074 the shape of $p(z)$ for the majority of the test sample, where
 1075 photometric errors on the faint galaxies lead to somewhat
 1076 broad peaks in the redshift posterior. However, as can be
 1077 seen in e. g. the top left panel of Fig. 1, for bright galaxies
 1078 with narrow $p(z)$ the grid spacing of $\delta z = 0.01$ is not suffi-
 1079 cient to resolve the peak. This is consistent with the results
 1080 described in Malz et al. (2018), who find that quantiles (and,
 1081 to a lesser degree, samples) often outperform gridded $p(z)$,
 1082 particularly for bright objects and in the presence of harsher
 1083 storage constraints. With a full 200 numbers to capture the

1084 information of each photo-z PDF, any parametrization will
 1085 perform adequately, but other storage parametrizations and
 1086 limits on storage resources may be considered in future work.
 1087 We will discuss this further in Section 6.

1088 Fig. 2 shows both the quantile-quantile plots (red) and
 1089 the histogram of PIT values (blue) summarizing the results
 1090 from each photo-z code. The red line shows the measured
 1091 quantiles, while the black diagonal represents the ideal QQ
 1092 values if the distribution were perfectly reproduced. A sec-
 1093 ond panel below the main panel for each code shows the dif-
 1094 ference between Q_{data} and Q_{theory} , i. e. the departure from
 1095 the diagonal, for clarity. Biases and trends in whether the
 1096 average width of the $p(z)$ values being over/under-predicted
 1097 are evident. An overall bias where the predicted redshift
 1098 is systematically low manifests as the measured QQ value
 1099 falling above the diagonal, as is the case for BPZ and EAZY,
 1100 while a systematic overprediction shows up as the measured
 1101 QQ value falling below the diagonal, as seen in TPZ. In
 1102 terms of PIT histograms, a systematic underprediction of
 1103 redshift corresponds to fewer PIT values at $PIT < 0.5$ and
 1104 more at $PIT > 0.5$, while a systematic overprediction will
 1105 show the opposite.

1106 Examination of the PIT histograms and QQ plots shows
 1107 that there are fairly generic issues with the width of $p(z)$ un-
 1108 certainties: DELIGHT, NN, SKYNET and TPZ all show a PIT
 1109 histogram with an dearth of low values and an excess of
 1110 high values, signs that, on average, their $p(z)$ are more
 1111 broad than the true distribution of redshifts. METAPHOR
 1112 shows the opposite trend, indicating the $p(z)$ are more
 1113 narrow than the distributions given by the true redshifts.
 1114 In all of these code cases there is a free parameter or band-
 1115 width that can be used to tune uncertainties. The sensitivity
 1116 of multiple codes to this bandwidth choice emphasizes the
 1117 fact that great care must be taken in setting user-defined
 1118 parameters in photo-z codes, even in the presence of rep-
 1119 resentative training/validation data. for FLEXZBOOST the
 1120 “sharpening” parameter (described in Section 3.2.4) plays
 1121 a key role in improving the results, resulting in a QQ plot
 1122 that is very nearly diagonal. A similar sharpening procedure
 1123 could be beneficial for several codes. Interestingly, the three
 1124 purely template-based codes, BPZ, EAZY, and LEPHARE,
 1125 show relatively well behaved $p(z)$ statistics (albeit with some
 1126 bias), which may indicate that the likelihood estimation with
 1127 representative templates is accurately capturing the uncer-
 1128 tainties on individual redshifts.

1129 The ideal PIT histogram would follow the black dashed
 1130 line, representing a uniform distribution of PIT values,
 1131 equivalent to the diagonal line in the QQ plot. Overly broad
 1132 $p(z)$ values show up as an excess of PIT values near 0.5
 1133 and a dearth of values at the edges, while overly narrow
 1134 $p(z)$ will have an excess at the edges and will be missing
 1135 values at the centre. Another feature evident in the PIT
 1136 histograms is the number of “catastrophic outlier” values
 1137 where the true redshift falls outside of the non-zero support
 1138 of $p(z)$, corresponding to $PIT = 0.0$ or 1.0 is more apparent
 1139 than in the QQ plots. Following Kodra & Newman (in prep.)
 1140 we define f_0 as the fraction of objects with $PIT < 0.0001$
 1141 or $PIT > 0.9999$. Table 2 lists these fractions for each of
 1142 the codes. For a proper Uniform distribution we expect a
 1143 value of 0.0002. Several codes show a marked excess, with
 1144 ANNz2, FLEXZBOOST, LEPHARE, AND METAPHOR with
 1145 $f_0 > 0.02$, indicating a sizeable number of catastrophic red-

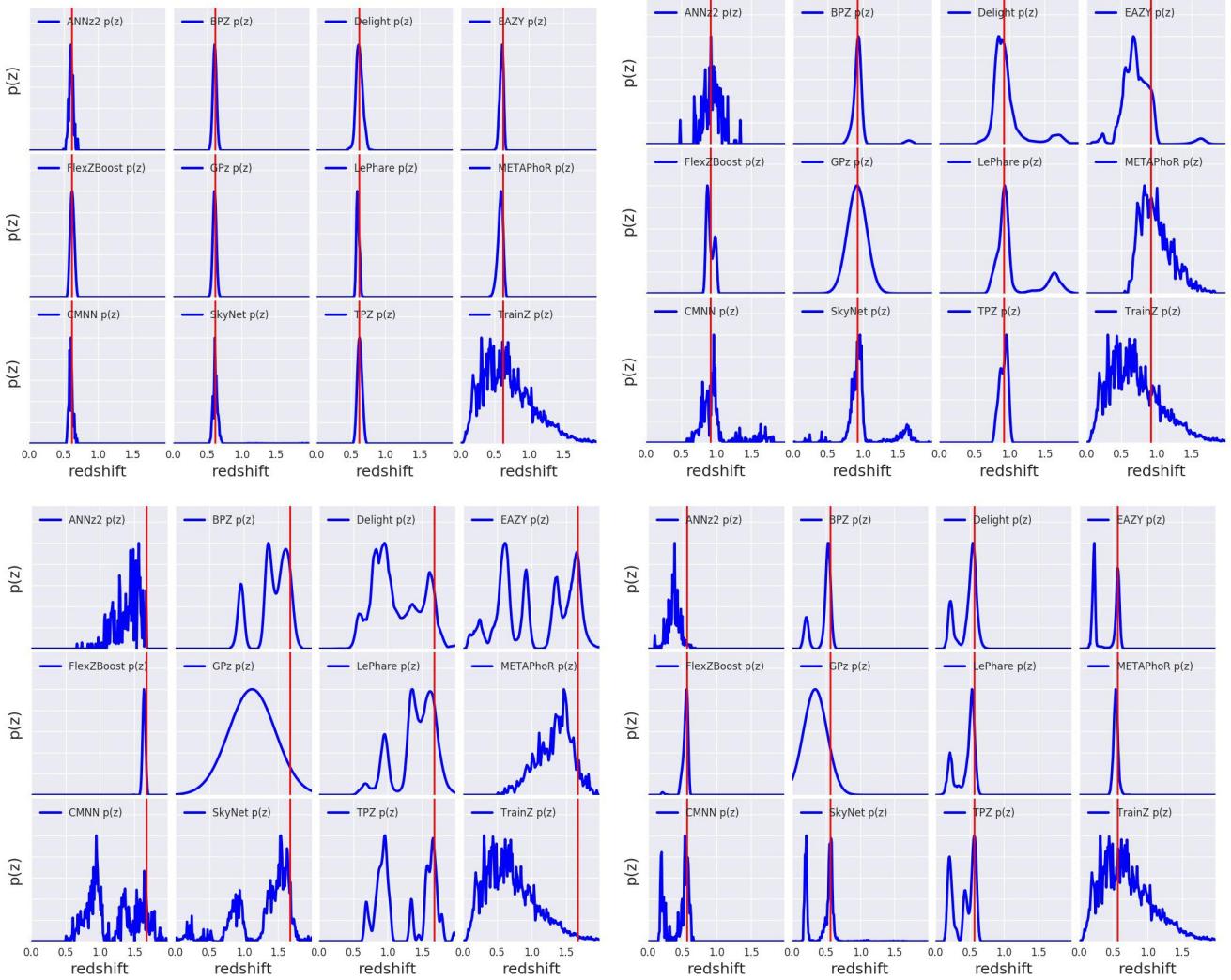


Figure 1. Four illustrative examples of individual $p(z)$ distributions produced by the codes. The red vertical line represents the true redshift. Examples are chosen with common features seen in PDFs: tight unimodal $p(z)$ (upper left), broad unimodal $p(z)$ (upper right), bimodal $p(z)$ (lower right), 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.

shift solutions where the true redshift is not covered by the extent of $p(z)$. For METAPHOR this may be partially due to an overall underprediction of the $p(z)$ width, however this is not the case for the other codes. LEPHARE is a particular outlier with nearly 5 per cent of objects outside of $p(z)$ support. Further study will be necessary to determine what is causing these misclassifications for LEPHARE. As expected, and by design, TRAINZ has the proper fraction of outliers for the f_0 statistic.

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

distribution. We calculate the AD statistic over the range of PIT values $v = [0.01, 0.99]$. ANNz2 and FLEXZBOOST score very well for the PIT metrics. METAPHOR and LEPHARE score very well in the PIT AD statistic, but both have a large number of catastrophic outliers, resulting in higher KS and CvM scores.

Given the near-perfect training data, examining the individual codes for explanations for departures from the expected behaviour will be instructive in avoiding similar problems in future tests. ANNz2 performs quite well in $p(z)$ based metrics. In the specific implementation employed in this paper, the final $p(z)$ is a weighted average of five neural-nets. During the training process ANNz2 compares the percentiles of the redshift training sample against the CDFs of the $p(z)$ sample. Distributions that more closely match are given extra weight, and the final weights are designed to produce accurate percentiles. Given that our metrics are focused on the percentile distributions, it is unsurprising that

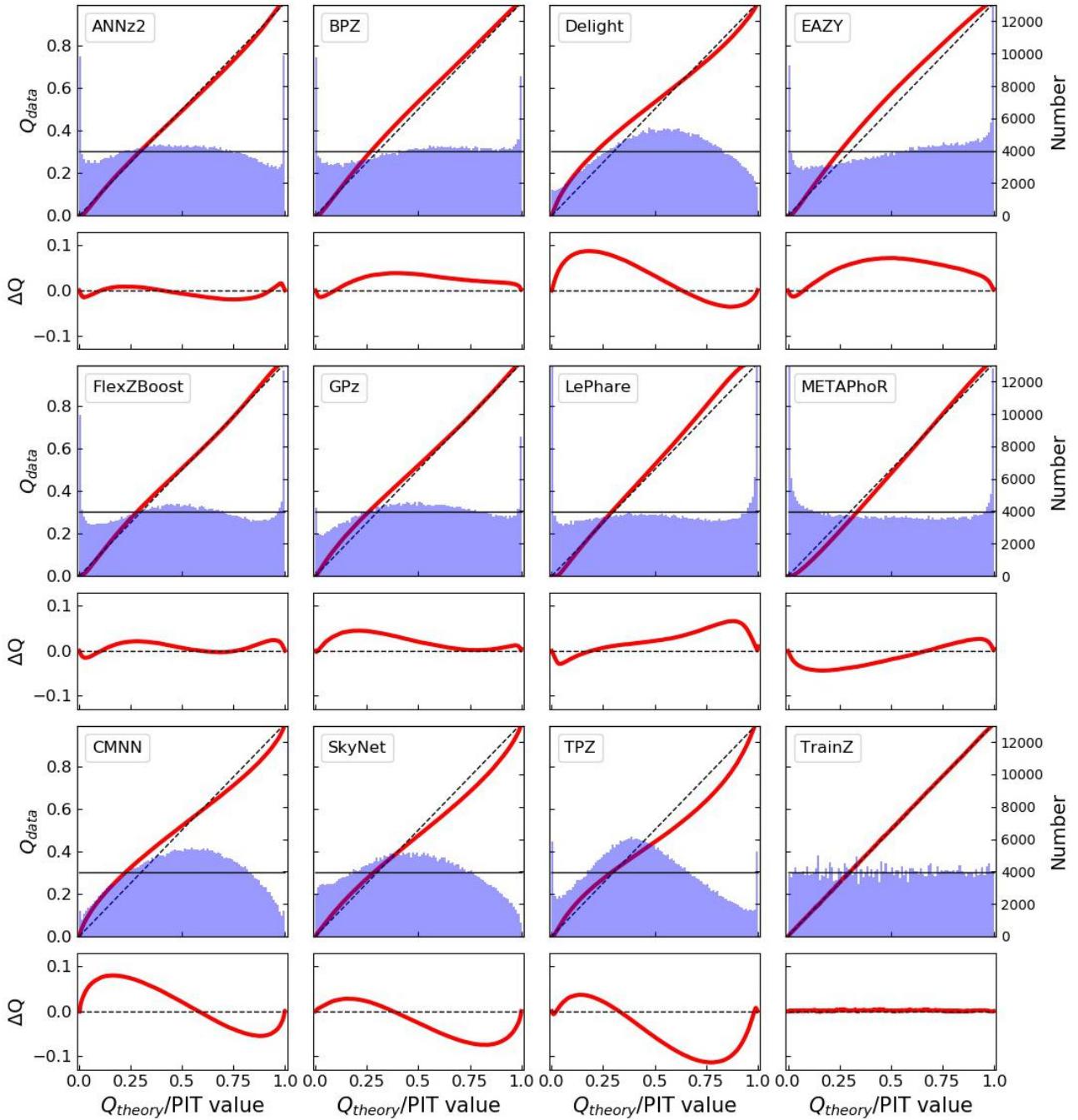


Figure 2. Summary plots for all eleven photo-z codes illustrating performance for the interim posterior statistics. The top panel of each pair shows both the Quantile-Quantile (QQ) plot (red) and the histogram of PIT values (blue). The desired behavior is a QQ plot that matches the diagonal dashed line, 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.

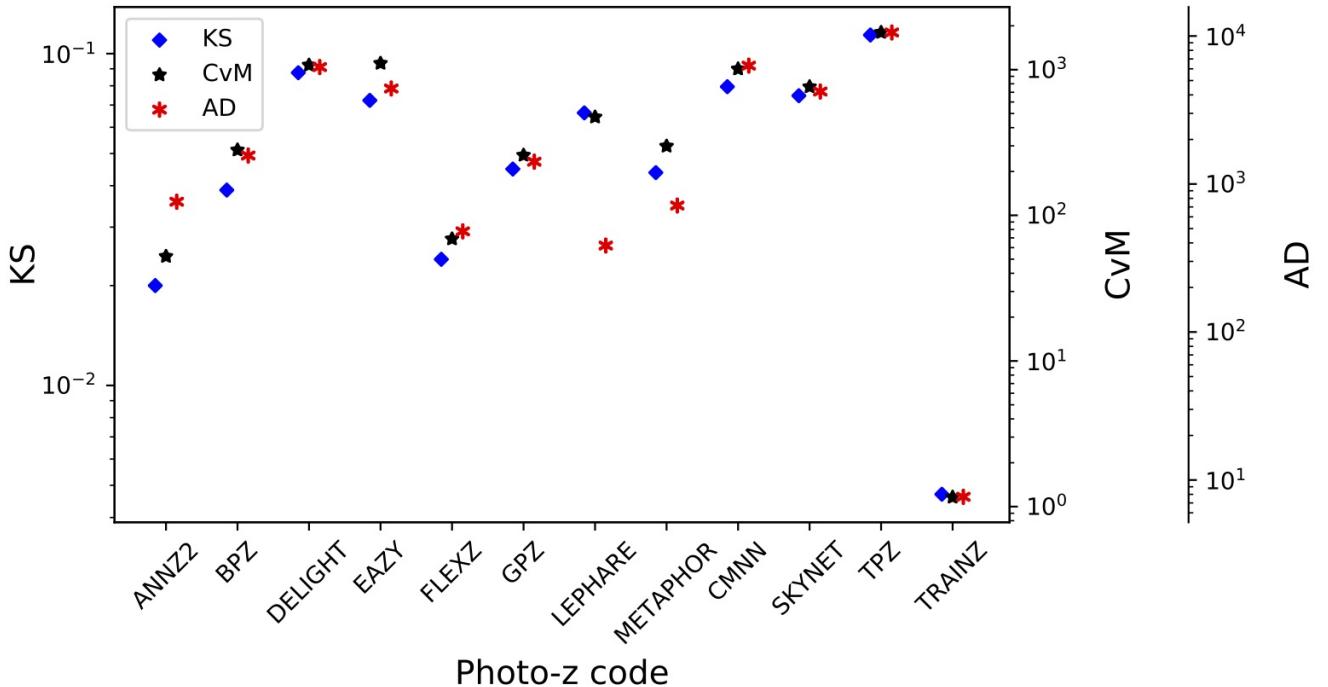


Figure 3. A visual representation of the Kolmogorov-Smirnov (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the PIT distributions. The statistics are often highly correlated, though the AD statistic truncates the extrema of the distribution 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	0.0068
LEPHARE	0.0486
METAPHOR	0.0229
CMNN	0.0034
SKYNET	0.0001
TPZ	0.0130
TRAINZ	0.0002

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 display $N'(z)$ 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. Several of the machine learning based codes appear to be overtrained, 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

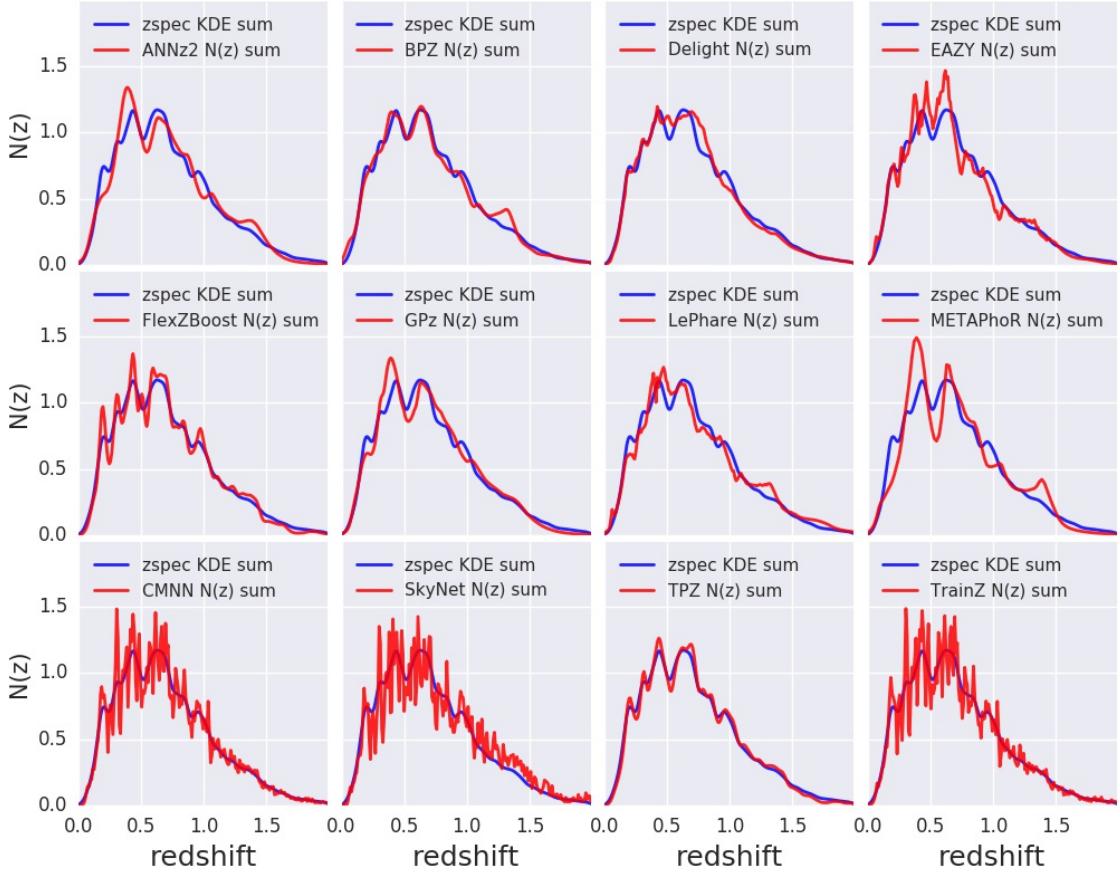


Figure 4. The stacked $p(z)$ produced by each photo- z code ($\hat{N}(z)$, red) compared to the spectroscopic redshift distribution ($N'(z)$, blue). Varying levels of small-scale structure are seen in the codes. $N'(z)$ is smoothed using a single bandwidth chosen via Scott’s rule for all codes.

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 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

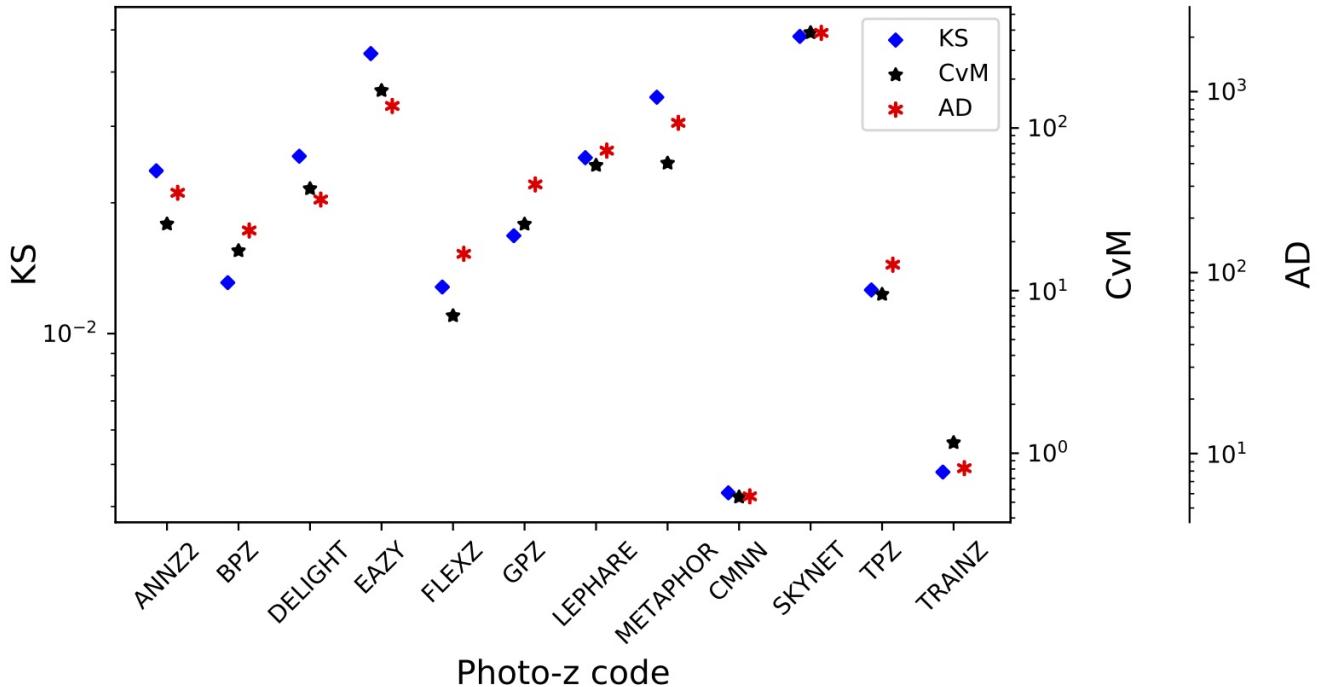


Figure 5. A visual representation of the Kolmogorov-Smirnov (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the $\hat{N}(z)$ distributions. The statistics are correlated, the codes with the lowest KS statistics tend to have the lowest CvM 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.

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 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 ANN22 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.

5.3 Interpretation of metrics

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

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

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

In summary, context is crucial to interpreting metrics and defending against the likes of TRAINZ. The best photo- z method is the one that most effectively achieves our science goals, not the one that performs best on a metric that does not accurately reflect those goals. In the absence of clear goals or the information necessary for a principled metric

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	-9.56
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	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	0.048	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

definition, we must think carefully before choosing a single metric

6 DISCUSSION

In this paper we presented results evaluating the photometric redshift PDF computation for eleven photo- z codes. As discussed in Section 4 the $p(z)$ should accurately reflect the relative likelihood as a function of redshift for each galaxy. All codes were provided a set of representative training data and tested on an idealized set of model galaxies with high signal-to-noise and photometry with no confounding effects due to blending, instrumental effects, the night sky, etc... included. The goal was not to determine a “best” photo- z code: in many ways, this was a baseline test of a “best case scenario” to predict the expected photo- z performance if a stage IV dark energy survey was to obtain complete training samples and perfectly calibrated their multi-band photometry. Given these idealized conditions, any deficiencies observed in a photo- z code’s performance should be a cause for concern, and may be evidence 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 prob-

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	0.692	0.605	0.619
LEPHARE	0.718	0.668	0.741
METAPHO.R	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

lems 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 $\hat{N}(z)$ the resulting metrics would be different for most, if not all, of the codes, as they would optimize toward a different goal. Specific metric choice can affect which codes are among the “best” codes. As stated earlier, there are cosmological science cases that require either individual galaxy photo-z measures, or ensemble $\hat{N}(z)$ measures. We must be aware of that the optimal method for one is not necessarily optimal for the other, and in fact several photo-z algorithms may be necessary in the final cosmological analysis in order to satisfy the requirements of all science use cases. The example of the simple TRAINZ estimator described in Section 5.3 shows a simple model with a $p(z)$ that is unrealistic for individual objects can still score very well on many of our metrics. It is important to look at all metrics, and keep in mind what

information each metric conveys. We re-emphasize that the dataset tested was quite idealized, and discuss enhancements that will be added in future simulations to test photo-z codes on increasingly realistic conditions in the following section.

6.1 Future work

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

1504 to identify AGN dominated galaxies, but also obtain more accurate
 1505 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 thousands 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.

1552 7 CONCLUSION

1553 In this study we have not accounted for the presence of Active Galactic Nuclei (AGN) contributions to galaxy fluxes.
 1554 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 iden-

1624 **Acknowledgments**

1625 Author contributions are listed below.
 1626 *S.J. Schmidt:* Led the project. (conceptualization, data
 1627 curation, formal analysis, investigation, methodology,
 1628 project administration, resources, software, supervision,
 1629 visualization, writing – original draft, writing – review &
 1630 editing)
 1631 *A.I. Malz:* Contributed to choice of metrics, implementation
 1632 in code, and writing. (conceptualization, methodology,
 1633 project administration, resources, software, visualization,
 1634 writing – original draft, writing – review & editing)
 1635 *J.Y.H. Soo:* Ran ANNz2 and Delight, updated abstract,
 1636 edited sections 1 through 6, added tables in Methods
 1637 and Results, updated references.bib and added references
 1638 throughout the paper
 1639 *M. Brescia:* main ideator of METAPHOR and of MLPQNA;
 1640 modification of METAPHOR pipeline to fit the LSST data
 1641 structure and requirements
 1642 *S. Cavuoti:* Contributed to choice and test of metrics, ran
 1643 METAPHOR, minor text editing
 1644 *G. Longo:* Scientific advise, test and validation of the
 1645 modified METAPHOR pipeline, text of the METAPHOR
 1646 section
 1647 *I.A. Almosallam:* vetted the early versions of the data set
 1648 and ran many photo-z codes on it, applied GPz to the final
 1649 version and wrote the GPz subsection
 1650 *M.L. Graham:* Ran the colour-matched nearest-neighbours
 1651 photo-z code on the Buzzard catalog and wrote the relevant
 1652 piece of Section 2; participated in discussions of the analy-
 1653 sis.
 1654 *A.J. Connolly:* Developed the colour-matched nearest-
 1655 neighbours photo-z code; participated in discussions of the
 1656 analysis.
 1657 *E. Nourbakhsh:* Ran and optimized TPZ code on the
 1658 Buzzard catalog and wrote a subsection of Section 2 for that
 1659 *J. Cohen-Tanugi:* contributed to running code, analysis
 1660 discussion, and editing, reviewing the paper
 1661 *H. Tranin:* contributed to providing SkyNet results and
 1662 writing the relevant section
 1663 *P.E. Freeman:* Contributed to choice of CDE metrics and
 1664 to implementation of FlexZBoost
 1665 *K. Iyer:* assisted in writing metric functions used to evaluate
 1666 codes
 1667 *J.B. Kalmbach:* Worked on preparing the figures for the
 1668 paper.
 1669 *E. Kovacs:* Ran simulations, discussed data format and
 1670 properties for SEDs, dust, and ELG corrections
 1671 *A.B. Lee:* Co-developed FlexZBoost and the CDE loss statis-
 1672 tic, wrote text on the work, and supervised the development
 1673 of FlexZBoost software packages
 1674 *C. Morrison:* Managerial support; Discussions with authors
 1675 regarding metrics and style; Some coding contribution to
 1676 metric computation.
 1677 *J. Newman:* Contributions to overall strategy, design of
 1678 metrics, and supervision of work done by Rongpu Zhou
 1679 *E. Nuss:* contributed to running code, analysis discussion,
 1680 and editing, reviewing the paper
 1681 *T. Pospisil:* Co-developed FlexZBoost software and CDE
 1682 loss calculation code
 1683 *M.J. Jarvis:* Contributed text on AGN to Discussion section
 1684 and portions of GPz work

1685 *R. Izbicki:* Co-developed FlexZBoost and the CDE loss
 1686 statistic, and wrote software for FlexZBoost

1687 The authors would like to thank their LSST-DESC pub-
 1688 lication review committee.

1689 *AIM* is advised by David W. Hogg and was supported by
 1690 National Science Foundation grant AST-1517237.

1691 The DESC acknowledges ongoing support from the In-
 1692 stitut National de Physique Nucléaire et de Physique des
 1693 Particules in France; the Science & Technology Facilities
 1694 Council in the United Kingdom; and the Department of En-
 1695 ergy, the National Science Foundation, and the LSST Cor-
 1696 poration in the United States. DESC uses resources of the
 1697 IN2P3 Computing Center (CC-IN2P3-Lyon/Villeurbanne -
 1698 France) funded by the Centre National de la Recherche Sci-
 1699 entifique; the National Energy Research Scientific Com-
 1700 puting Center, a DOE Office of Science User Facility supported
 1701 by the Office of Science of the U.S. Department of Energy
 1702 under Contract No. DE-AC02-05CH11231; STFC DiRAC
 1703 HPC Facilities, funded by UK BIS National E-infrastructure
 1704 capital grants; and the UK particle physics grid, supported
 1705 by the GridPP Collaboration. This work was performed in
 1706 part under DOE Contract DE-AC02-76SF00515.

APPENDIX A: POINT ESTIMATE PHOTOMETRIC REDSHIFTS

1707 While we do not recommend the use of single point estimates
 1708 of redshift for most science applications, plots of the point
 1709 estimates can be a useful qualitative diagnostic of photo-z
 1710 code performance, i. e. examining point photo-z vs. spec-z
 1711 plots visually can give a quick impression of some common
 1712 trends in different codes. Computing point estimate statistics
 1713 may also be useful for more direct comparisons with previous
 1714 photo-z evaluations. If a point-estimate is preferred for a
 1715 specific science case, it is fairly simple to compute the mean,
 1716 mode, or some other simple estimator from each $p(z)$, so
 1717 these point estimates can be easily derived from the stored
 1718 $p(z)$.

1719 There are several common point estimators of photo-z
 1720 posteriors employed by different codes, e. g. the mode, mean,
 1721 median of the $p(z)$ distribution. In addition, many of the
 1722 machine learning based estimators can be set up to return a
 1723 single redshift solution. For example, SkyNet can be config-
 1724 ured to run as a regressor that returns a single float rather
 1725 than a classifier that returns a 200-bin $p(z)$ estimate. The
 1726 single value returned by a machine learning based code may
 1727 not correspond to a particular measure such as the mode or
 1728 mean, and so to avoid interpretation of results that might be
 1729 introduced by variations in choice of specific point-estimate
 1730 implementation per code, we discard the code-specific point
 1731 estimates. We instead calculate point estimates more uni-
 1732 formly across the codes directly from the $p(z)$ using two mea-
 1733 sures, z_{PEAK} and z_{WEIGHT} . z_{PEAK} is simply the maximum
 1734 value attained for each galaxy $p(z)$, the mode of the proba-
 1735 bility distribution. z_{WEIGHT} is defined similarly to how it
 1736 is defined in Dahlen et al. (2013), as the weighted mean of
 1737 the redshift over the main peak of $p(z)$ containing the z_{PEAK}
 1738 value. The main peak is defined by subtracting $0.05 \times z_{\text{PEAK}}$
 1739 from $p(z)$ and identifying the roots to isolate the peak con-
 1740 taining z_{PEAK} , z_{WEIGHT} is defined as the weighted mean

redshift within this peak. We restrict to a single peak in order to avoid confusion from bimodal and multimodal $p(z)$ such as those shown in bottom panels of Figure 1. For example, for a bimodal probability distribution a weighted mean calculated over both peaks would fall between the peaks, at a redshift where the probability is minimal. Restricting the weighting to a single peak ensures that the point estimate will fall in the region of maximum redshift probability.

A1 Point Estimate Metrics

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

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

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

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

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

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

weighting of z_{WEIGHT} , resulting in a smoother distribution of redshift estimates. Several features perpendicular to the main $z_{phot} = z_{spec}$ line are evident. These features are due to the 4000 angstrom break passing through the gaps between adjacent LSST filters. These features are most prominent in template-based codes, but appear to some degree in all codes tested.

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

Finally, we note that all eleven codes are at or near the goals for point-estimates as outlined in the LSST Science Requirements Document¹⁹ and Graham et al. (2018). This is to be expected, given that the requirements were designed such that a point estimate photo-z would meet these requirements for perfect training data to a depth of $i < 25$. But, it is still an encouraging sign, given an updated mock galaxy simulation and the expanded set of photo-z codes tested.

REFERENCES

- Abbott T., et al., 2005, preprint (arXiv:astro-ph/0510346)
 Abell P. A., et al., 2009, preprint (arXiv:0912.0201),
 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
 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, California, U.S.A
 Brescia M., Caruoti S., Amaro V., Riccio G., Angora G., Vellucci C., Longo G., 2018, preprint, (arXiv:1802.07683)
 Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, **432**, 1483
 Carrasco Kind M., Brunner R. J., 2014, *MNRAS*, **442**, 3380
 Caruoti S., Amaro V., Brescia M., Vellucci C., Tortora C., Longo G., 2017, *MNRAS*, **465**, 1959
 Chen T., Guestrin C., 2016, in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD ’16. ACM, New York, NY, USA, pp 785–794, doi:10.1145/2939672.2939785, <http://doi.acm.org/10.1145/2939672.2939785>
 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

¹⁹ available at: <http://ls.st/srd>

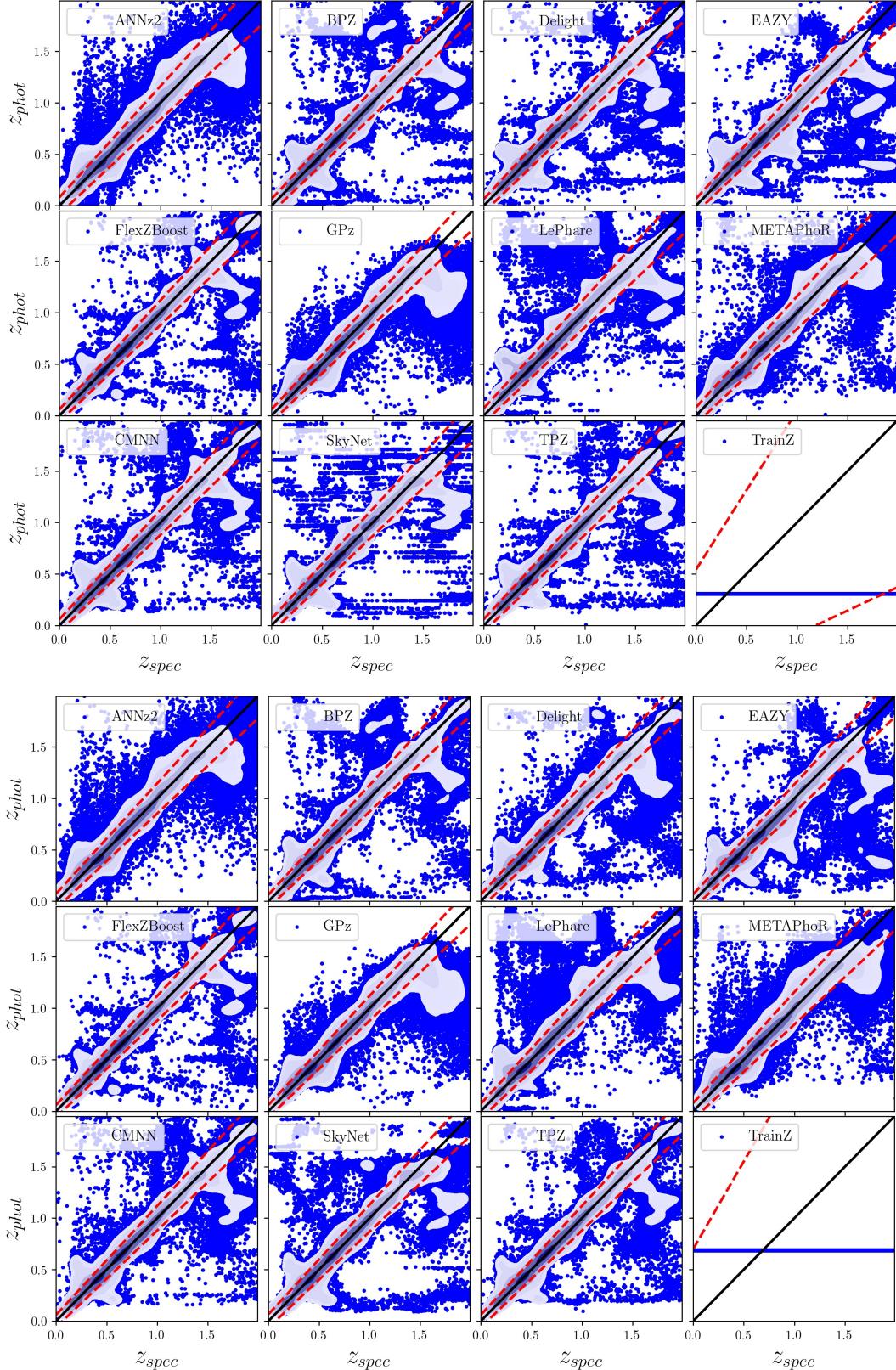


Figure A1. Point estimate photo-z's derived from the posteriors. Top panel shows z_{PEAK} , while bottom panel shows z_{WEIGHT} . Point estimate density is represented with fixed density contours, while outliers at lower density are represented by blue points. While use of point-estimate photo-z's is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

Table A1. Point estimate statistics

Photo-z Code	<i>Z_{PEAK}</i>			<i>Z_{WEIGHT}</i>		
	$\frac{\sigma_{IQR}}{(1+z)}$	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	0.0202	-0.00091	0.036	0.0201	-0.000950	0.037
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

- 1860 Fernández-Soto A., Lanzetta K. M., Yahil A., 1999, *ApJ*, **513**, 34
1861 Firth A. E., Lahav O., Somerville R. S., 2003, *MNRAS*, **339**, 1195
1862 Freeman P. E., Izbicki R., Lee A. B., 2017, *MNRAS*, **468**, 4556
1863 Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, *MNRAS*, 441, 1741
1864 Graham M. L., Connolly A. J., Ivezić Ž., Schmidt S. J., Jones R. L., Jurić M., Daniel S. F., Yoachim P., 2018, *AJ*, **155**, 1
1865 Green J., et al., 2012, preprint (arXiv:1208.4012),
1866 Hildebrandt H., et al., 2010, *A&A*, **523**, A31
1867 Ilbert O., et al., 2006, *A&A*, **457**, 841
1868 Ivezić Ž., et al., 2008, preprint (arXiv:0805.2366),
1869 Izbicki R., Lee A. B., 2017, *Electron. J. Statist.*, **11**, 2800
1870 Izbicki R., Lee A. B., Freeman P. E., 2017, *Ann. Appl. Stat.*, **11**, 698
1871 Laureijs R., et al., 2011, preprint (1110.3193),
1872 Leistedt B., Hogg D. W., 2017, *ApJ*, **838**, 5
1873 Maddox N., Hewett P. C., Warren S. J., Croom S. M., 2008,
1874 *MNRAS*, **386**, 1605
1875 Malz A., Hogg D., in prep., CHIPPR, chippr
1876 Malz A., Marshall P., DeRose J., Graham M., Schmidt S., Wechsler R., 2018, *AJ*, Accepted,
1877 Mandelbaum R., et al., 2008, *MNRAS*, **386**, 781
1878 Massarotti M., Iovino A., Buzzoni A., 2001, *A&A*, **368**, 74
1879 Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes J. D., Castander F. J., Paltani S., 2017, *ApJ*, **841**, 111
1880 Newman J. A., 2008, *ApJ*, **684**, 88
1881 Newman J. A., et al., 2015, *Astroparticle Physics*, **63**, 81
1882 Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008,
1883 *ApJ*, **689**, 709
1884 Polsterer K. L., D'Isanto A., Gieseke F., 2016, preprint
1885 (arXiv:1608.08016),
1886 Rasmussen C., Williams C., 2006, *Gaussian Processes for Machine Learning. Adaptative computation and machine learning series*, MIT Press, Cambridge, MA
1887 Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, *ApJ*, **771**, 30
1888 Richards G. T., et al., 2006, *ApJS*, **166**, 470
1889 Richards G. T., et al., 2015, *ApJS*, **219**, 39
1890 Sadeh I., Abdalla F. B., Lahav O., 2016, *PASP*, **128**, 104502
1891 Salvato M., et al., 2011, *ApJ*, **742**, 61
1892 Sánchez C., et al., 2014, *MNRAS*, **445**, 1482
1893 Scott D. W., 1992, *Multivariate Density Estimation. Theory, Practice, and Visualization*. Wiley
1894 Skrutskie M. F., et al., 2006, *AJ*, **131**, 1163
1895 Tanaka M., et al., 2017, preprint (arXiv:1704.05988),
1896 York D. G., et al., 2000, *AJ*, **120**, 1579
1897 de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, *Exp. Astron.*, **35**, 25
1898 de Jong J. T. A., et al., 2017, *A&A*, **604**, A134