MNRAS **000**, 000–000 (0000) Preprint 20 November 2019 Compiled using MNRAS LaTeX style file v3.0

Evaluation of probabilistic photometric redshift estimation approaches for LSST

S.J. Schmidt[1], A.I. Malz[2,3,4], J.Y.H. Soo[5], I.A. Almosallam[6,7], M. Brescia[8], S. Cavuoti[8,9], J. Cohen-Tanugi[10], A.J. Connolly[11], P.E. Freeman[12], M.L. Graham[11], K. Iyer[13], M.J. Jarvis[14,15], J.B. Kalmbach[16], E. Kovacs[17], A.B. Lee[12], G. Longo[9], C. B. Morrison[11], J. Newman[18], E. Nourbakhsh[19], E. Nuss[10], T. Pospisil[12], H. Tranin[10], R. Zhou[18], R. Izbicki[20,21]

(LSST Dark Energy Science Collaboration)

[1] *Department of Physics, University of California, One Shields Ave., Davis, CA, 95616, USA*
[2] *German Centre of Cosmological Lensing, Ruhr-Universitaet Bochum, Universitaetsstraße 150, 44801 Bochum, Germany*
[3] *Center for Cosmology and Particle Physics, New York University, 726 Broadway, New York, 10003, USA*
[4] *Department of Physics, New York University, 726 Broadway, New York, 10003, USA*
[5] *Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[6] *King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia*
[7] *Information Engineering, Parks Road, Oxford, OX1 3PJ, UK*
[8] *INAF-Capodimonte Observatory, Salita Moiariello 16, I-80131, Napoli, Italy*
[9] *Department of Physics E. Pancini, University Federico II, via Cinthia 6, I-80126, Napoli, Italy*
[10] *Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS, Montpellier, France*
[11] *Department of Astronomy, University of Washington, Box 351580, U.W., Seattle WA 98195, USA*
[12] *Department of Statistics & Data Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*
[13] *Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, NJ 08854-8019 USA*
[14] *Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK*
[15] *Department of Physics and Astronomy, University of the Western Cape, Bellville 7535, South Africa*
[16] *Department of Physics, University of Washington, Box 351560, Seattle, WA 98195, USA*
[17] *Argonne National Laboratory, Lemont, IL 60439, USA*
[18] *Department of Physics and Astronomy and the Pittsburgh Particle Physics, Astrophysics and Cosmology Center (PITT PACC), University of Pittsburgh, Pittsburgh, PA 15260, USA*
[19] *Deptartment of Physics, University of California, One Shields Ave., Davis, CA, 95616, USA*
[20] *Department of Statistics, Federal University of Sao Carlos, Sao Carlos, Brazil*
[21] *External collaborator*

20 November 2019

ABSTRACT

Many scientific investigations of photometric galaxy surveys require redshift estimates, whose uncertainty properties are best encapsulated by photometric redshift (photo-z) posterior probability distribution functions (PDFs). A plethora of photo-z PDF estimation methodologies abound, producing discrepant results with no consensus on a preferred approach. We present the results of a comprehensive experiment comparing twelve photo-z algorithms applied to mock data produced for the Large Synoptic Survey Telescope (LSST) Dark Energy Science Collaboration (DESC). By supplying perfect prior information, in the form of the complete template library and a representative training set as inputs to each code, we demonstrate the impact of the assumptions underlying each technique on the output photo-z PDFs. In the absence of a notion of true, unbiased photo-z PDFs, we evaluate and interpret multiple metrics of the ensemble properties of the derived photo-z PDFs as well as traditional reductions to photo-z point estimates. We report systematic biases and overall over/under-breadth of the photo-z PDFs of many popular codes, which may indicate avenues for improvement in the algorithms or implementations. Furthermore, we raise attention to the limitations of established metrics for assessing photo-z PDF accuracy; though we identify the conditional density estimate (CDE) loss as a promising metric of photo-z PDF performance in the case where true redshifts are available but true photo-z PDFs are not, we emphasize the need for science-specific performance metrics.

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

1 INTRODUCTION

The current and next generations of large-scale galaxy surveys, including the Dark Energy Survey (DES, Abbott et al. 2005), the Kilo-Degree Survey (KiDS, de Jong et al. 2013), Hyper Suprime-Cam Survey (HSC, Aihara et al. 2018a,b), Large Synoptic Survey Telescope (LSST, Abell et al. 2009), Euclid (Laureijs et al. 2011), and Wide-Field Infrared Survey Telescope (WFIRST, Green et al. 2012), represent a paradigm shift to reliance on photometric, rather than solely spectroscopic, galaxy catalogues of substantially larger size at a cost of lacking complete spectroscopically confirmed redshifts (z). Effective astrophysical inference using the catalogues resulting from these ongoing and upcoming missions, however, necessitates accurate and precise photometric redshift (photo-z) estimation methodologies.

As an example, in order for photo-z systematics to not dominate the statistical noise floor of LSST's main cosmological sample of several 10^9 galaxies, the LSST Science Requirements Document (SRD)[1] specifies that individual galaxy photo-zs must have root-mean-square error $\sigma_z < 0.02(1 + z)$, 3σ catastrophic outlier rate below 10 per cent, and bias below 0.003. Specific science cases may have their own requirements on photo-z performance that exceed those of the survey as a whole. In that vein, the LSST Dark Energy Science Collaboration (LSST-DESC) developed a separate SRD (The LSST Dark Energy Science Collaboration et al. 2018) that conservatively forecasts the constraining power of five cosmological probes, leading to even more stringent requirements on photo-z performance, including those defined in terms of tomographically binned subsample populations rather than individual galaxies.

Though the standard has long been for each galaxy in a photometric catalogue to have a photo-z point estimate and Gaussian error bar, even early applications of photo-zs in precision cosmology indicate the inadequacy of point estimates (Mandelbaum et al. 2008) to encapsulate the degeneracies resulting from the nontrivial mapping between broad band fluxes and redshift. Far from a hypothetical situation, such degeneracies are real consequences of the same deep imaging that enables larger galaxy catalogue sizes. The lower luminosity and higher redshift populations captured by deeper imaging introduce major physical systematics to photo-zs, among them the Lyman break/Balmer break degeneracy, that did not affect shallower large area surveys like the Sloan Digital Sky Survey (SDSS, York et al. 2000) and Two Micron All Sky Survey (2MASS, Skrutskie et al. 2006).

To fully characterize such physical degeneracies, subsequent photometric galaxy catalogue data releases, (e. g. Sheldon et al. 2012; Erben et al. 2013; de Jong et al. 2017), provide a more informative photo-z data product, the photo-z probability density function (PDF), that describes the redshift probability, commonly denoted as $p(z)$, as a function of a galaxy's redshift, conditioned on the observed photometry. Early template-based methods such as Fernández-Soto et al. (1999) approximated the likelihood of photometry conditioned on redshift with the relative χ^2 values of template spectra. Not long after, Bayesian adaptations of template-based approaches such as Benítez (2000) combined the estimated likelihoods with a prior to yield a posterior PDF of redshift conditioned on photometry. While the first data-driven photo-z algorithms yielded a point estimate, Firth et al. (2003) estimated a photo-z PDF using a neural net with realizations scattered within the photometric errors.

There are numerous techniques for deriving photo-z PDFs, yet no one method has been established as clearly superior. Consistent experimental conditions enable the quantification if not isolation of their differences, which can be interpreted as a sort of *implicit prior* imparted by the method itself. Comprehensive comparisons of photo-z methods have been made before; the Photo-z Accuracy And Testing (PHAT, Hildebrandt et al. 2010) effort focused on photo-z point estimates derived from many photometric bands. Rau et al. (2015) introduced a new method for improving photo-z PDFs using an ordinal classification algorithm. DES compared several codes for photo-z point estimates and a subset with photo-z PDF information (Sánchez et al. 2014) and examined summary statistics of photo-z PDFs for tomographically binned galaxy subsamples (Bonnett et al. 2016).

This paper is distinguished from other comparisons of photo-z methods by its focus on the evaluation criteria for photo-z PDFs and interpretation thereof. In the absence of simulated data drawn from known redshift distributions, the very concept of a "true PDF" for an individual galaxy is unavailable, and we must instead rely on measures of ensemble behaviour to characterize PDF quality (see § 4 for further discussion). We aim to perform a comprehensive sensitivity analysis of photo-z PDF techniques in order to ultimately select those that will become part of the LSST-DESC pipelines, described in the Science Roadmap (SRM)[2]. In this initial study, we focus on evaluating the performance of photo-z PDF codes using PDF-specific performance metrics in a formally controlled experiment with complete and representative prior information (template libraries and training sets) to set a baseline for subsequent investigations. This approach probes how each code considered exploits the information content of the data versus prior information from template libraries and training sets.

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

2 DATA

In order to test the current generation of photo-z PDF codes, we employ an existing simulated galaxy catalogue, described in detail in Section 2.1. The experimental conditions shared among all codes are motivated by the LSST SRD requirements and implemented for machine learning and template-

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

[2] Available at: `https://lsstdesc.org/assets/pdf/docs/DESC_SRM_latest.pdf`

based photo-z PDF codes according to the procedures of Sections 2.3.1 and 2.3.2 respectively.

2.1 The Buzzard-v1.0 simulation

Our mock catalogue is derived from the BUZZARD-highres-v1.0 catalogue (DeRose et al. 2019, Wechsler et al., in prep). BUZZARD is built on a dark matter-only N-body simulation of 2048^3 particles in a 400 Mpc h^{-1} box. The lightcone was constructed from smoothing and interpolation between a set of time snapshots. Dark matter halos were identified using the `Rockstar` software package (Behroozi et al. 2013) and then populated with galaxies with stellar masses and absolute r-band magnitudes in the SDSS system determined using a sub-halo abundance matching model constrained to match both projected two-point galaxy clustering statistics and an observed conditional stellar mass function (Reddick et al. 2013).

To assign a spectrum to each galaxy, the Adding Density Dependent Spectral Energy Distributions (SEDs) procedure (`ADDSEDS`, deRose in prep.)[3] was used. `ADDSEDS` uses a sample of $\sim 5 \times 10^5$ galaxies from the magnitude-limited SDSS Data Release 6 Value Added Galaxy Catalogue (NYU-VAGC, Blanton et al. 2005) to train an empirical relation between absolute r-band magnitude, local galaxy density, and SED. Each SDSS spectrum is parameterized by five weights corresponding to a weighted sum of five basis SED components using the `k-correct` software package[4] (Blanton & Roweis 2007).

Correlations between SED and galaxy environment were included so as to preserve the colour-density relation of galaxy environments. The distance to the spatially projected fifth-nearest neighbour was used as a proxy for local density in the SDSS training sample. For each simulated galaxy, a galaxy with similar absolute r-band magnitude and local galaxy density was chosen from the training set, and that training galaxy's SED was assigned to the simulated galaxy.

2.1.1 Caveats

By necessity, BUZZARD does not contain all of the complicating factors present in real data, and here we discuss the most pertinent ways that these limitations affect our experiment. BUZZARD includes only galaxies, not stars nor AGN. The catalogue-based construction excludes image-level effects, such as deblending errors, photometric measurement issues, contamination from sky background (Zodiacal light, scattered light, etc.), lensing magnification, and Galactic reddening.

The BUZZARD SEDs are drawn from a set of $\sim 5 \times 10^5$ SEDs, which themselves are derived from a five-component linear combination fit to $\sim 5 \times 10^5$ SDSS galaxies; thus the sample contains only galaxies that resemble linear combinations of those for which SDSS obtained spectra, and there are necessarily duplicates. The linear combination SEDs also restrict the properties of the galaxy population to linear combinations of the properties corresponding to five basis templates, precluding the modeling of non-linear features such as the full range of emission line fluxes relative to the continuum. The only form of intrinsic dust reddening comes from what is already present in the five basis SEDs via the training set used to create the basis templates, and linear combinations thereof do not span the full range of realistic dust extinction observed in galaxy populations.

While these idealized conditions limit the realism of our mock data, they are irrelevant to the controlled experimental conditions of this study, if anything assuring that differentiation in the performance of the photo-z PDF codes is due to the inferential techniques rather than nuances in the data.

2.2 LSST-like mock observations

Given the SED, absolute r-band magnitude, and true redshift of each simulated galaxy, we computed apparent magnitudes in the six LSST filter passbands, *ugrizy*. We assigned magnitude errors in the six bands using the simple model of Ivezić et al. (2008), assuming achievement of the full 10-year depth, with a modification of fiducial LSST total numbers of 30-second visits for photometric error generation: we assume 60 visits in u-band, 80 visits in g-band, 180 visits in r-band, 180 visits in i-band, 160 visits in z-band, and 160 visits in y-band.

As a consequence of adding Gaussian-distributed photometric errors, 2.0 per cent of our galaxies exhibit a negative flux in one or more bands, the vast majority of which are in the u-band. We deem such negative fluxes *non-detections* and assign a placholder magnitude of 99.0 in the catalogue to indicate to the photo-z PDF codes that such galaxies would be "looked at but not seen" in multi-band forced photometry.

The full dataset thus covers 400 square degrees and contains 238 million galaxies of redshift $0 < z \leq 8.7$ down to $r = 29$. Systematic inconsistencies with galaxy colors at $z > 2$ were observed, so the catalogue was limited to $0 < z \leq 2.0$. To obtain a catalogue matching the LSST Gold Sample, we imposed a cut of $i < 25.3$, which gives a signal-to-noise ratio $\gtrsim 30$ for most galaxies. In order for statistical errors to be subdominant to the systematic errors we aim to probe, we further reduced the sample size to $< 10^7$ galaxies by isolating ~ 16.8 square degrees selected from five separate spatial regions of the simulation. We refer to this final set of galaxies as DC1, for the first LSST-DESC Data Challenge.

2.3 Shared prior information

For the purpose of performing a controlled experiment that compares photo-z PDF codes on equal footing as a baseline for a future sensitivity analysis, we take care to provide each with optimal prior information. Redshift estimation approaches built upon physical modeling and machine learning alike have a notion of prior information considered beyond the photometry of the data for which redshift is to be constrained: that information is derived from a template library for a model-based code and a training set for a data-driven code. In this initial study, we seek to set a baseline for a later comparison of the performance of photo-z PDF codes under incomplete and non-representative prior information that will propagate differently in the space of data-driven

[3] https://github.com/vipasu/addseds
[4] http://kcorrect.org

and model-based algorithms. However, for the baseline case of perfect prior information, physical modeling and machine learning codes can indeed be put on truly equal footing. We outline the equivalent ways of providing all codes perfect prior information below.

2.3.1 *Training and test set division*

Following the findings of Bernstein & Huterer (2010), Newman et al. (2015), and Masters et al. (2017) that only $\sim 10^4$ spectra are necessary to calibrate photo-zs to Stage IV requirements, we aimed to set aside a randomly selected training set of $3-5\times10^4$ galaxies, ~ 10 per cent of the full sample. After all cuts described above, we designated the *DC1 training set* of 44 404 galaxies for which observed photometry, true SEDs, and true redshifts would be provided to all codes and the blinded *DC1 test set* of 399 356 galaxies for which photometry alone would be provided to all codes and photo-z PDFs would be requested. The exact form of LSST photometric filter transmission curves were also considered public information that could be used by any code.

2.3.2 *Template library construction*

We aimed to provide template-fitting codes with complete yet manageable library of templates spanning the space of SEDs of the DC1 galaxies. We constructed $K = 100$ representative templates from the $\sim 5\times10^5$ SEDs of the SDSS DR6 NYU-VAGC by using the five-dimensional vectors of SED weight coefficients described above. After regularizing the SED weight coefficients $\in [0, 1]$, we ran a simple K-means clustering algorithm on the five-dimensional space of regularized SED weight coefficients of the SDSS galaxy sample. The resulting clusters were used to define Voronoi cells in the space of weight coefficients, with centre positions corresponding to weights for the `k-correct` SED components, yielding the 100 SEDs that comprise the *DC1 template set* provided to all template-based codes. We did not, however, exclude from consideration template-based codes that made modifications in their use of these templates due to architecture limitations (as opposed to knowledge of the experimental conditions that could artificially boost the code's apparent performance), with deviations noted in Section 3.

3 METHODS

Here we summarize the twelve photo-z PDF codes compared in this study, listed in Table 1, which include both established and emerging approaches in template fitting and machine learning. Though not exhaustive, this sample represents codes for which there was sufficient expertise within the LSST-DESC Photometric Redshifts Working Group. Some aspects of data treatment were left to the individual code runners, for example, whether/how to split the available data with known redshifts into separate training and validation sets.

Another key difference is the treatment of non-detections in one or more bands: some codes ignore incomplete bands, while others replace the value with either an estimate for the detection limit, the mean of other values in the training set, or another default value. There are varying conventions among machine learning-based codes for treatment of non-detections, and no one prescription dominates in the photo-z literature. However, we remind the reader that only 2.0 per cent of our sample has non-detections, almost exclusively in the u-band, and thus should not dominate the code performance differences.

The authors welcome interest from those outside LSST-DESC to have their codes assessed in future investigations that build upon this one.

We describe the algorithms and implementations of the model-based and data-driven codes in Sections 3.1 and 3.2 respectively, with a straw-person approach included in Section 3.3.

3.1 Template-based Approaches

We test three publicly available and commonly used template-based codes that share the standard physically motivated approach of calculating model fluxes for a set of template SEDs on a grid of redshift values and evaluating a χ^2 merit function using the observed and model fluxes:

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

where A is a normalization factor, $F^i_{\rm pred}(T, z)$ is the flux predicted for a template T at redshift z, $F^i_{\rm obs}$ is the observed flux in a given band i, $\sigma^i_{\rm obs}$ is the observed flux error, and N_{filt} is the total number of filters, in our case the six *ugrizy* LSST filters. The likelihood is a sum of observed flux error $\sigma^{\rm obs}_b$-weighted squared differences between the observed flux $F^{\rm obs}_b$ and the normalized predicted flux $F^{\rm mod}_b(T, z)$ in $N_{\rm filt}$ photometric filters b, which are the LSST *ugrizy* filters in this case. Specific implementation details of each code, e. g. prior form and implementation, are described below.

3.1.1 *BPZ*

Bayesian Photometric Redshift (`BPZ`, Benítez 2000) determines the likelihood $p(C|z, T)$ of a galaxy's observed colours C for a set of SED templates T at redshifts z. The `BPZ` likelihood is related to the χ^2 likelihood by $p(C|z, T) \propto \exp[-\chi^2/2]$. Given a Bayesian prior $p(z, T|m_0)$ over apparent magnitude m_0 and type T, and assuming that the SED templates are spanning and exclusive, `BPZ` constructs the redshift posterior $p(z|C, m_0)$ by marginalizing over all SED templates with the form $p(z|C, m_0) \propto \sum_T p(C|z, T)\, p(z, T|m_0)$ (Eq. 3 from Benítez 2000), corresponding to setting the parameter `PROBS_LITE=TRUE` in the `BPZ` parameter file. The `BPZ` prior is the product of an SED template proportion that varies with apparent magnitude $p(T|m_0)$ and a prior $p(z|T, m_0)$ over the expected redshift as a function of apparent magnitude and SED template. We anticipate `BPZ` to outperform other template-based approaches due to the prior that both comprehensively accounts for SED type and is calibrated to the training set.

Here we test `BPZ`-v 1.99.3 (Benítez 2000) with the DC1 template set of Section 2.3.2. To keep the number of free parameters manageable, the DC1 template set is pre-sorted by the rest-frame $u-g$ colour and split into three broad classes of SED template, equivalent to the E, Sp and Im/SB types. The Bayesian prior term $p(T|m_0)$ was derived directly from

Table 1. List of photo-z PDF codes featured in this study

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

the DC1 training set, and the other term $p(z|T, m_0)$ was chosen to be the best fit for the eleven free parameters from the functional form of Benítez (2000). We use template interpolation, creating two linearly interpolated templates between each basis SED (sorted by rest-frame $u-g$ colour) by setting the parameter `INTERP=2`. Prior to running the code, the non-detection placeholder magnitude was replaced with an estimate of the 1-σ detection limit for the undetected band as a proxy for a value close to the estimated sky noise threshold.

3.1.2 *EAZY*

Easy and Accurate Photometric Redshifts from Yale (`EAZY`, Brammer et al. 2008) extends the basic χ^2 fit procedure that defines template-fitting approaches. The algorithm models the observed photometry with a linear combination of template SEDs at each redshift. The best-fit SED at each redshift is found by simultaneously fitting one, two, or all of the templates via χ^2 minimization, which is distinct from marginalizing across all templates. The minimized χ^2 likelihood at each redshift is then combined with an apparent magnitude prior to obtain the redshift posterior PDF. We note that the utilization of the best-fit SED conditioned on redshift rather than a proper marginalization does not lead to the correct posterior distribution, an implementation issue that has now been identified and will be addressed by the developers in the future.

In contrast with `BPZ`, `EAZY`'s apparent magnitude prior is independent of SED, though it was derived empirically from the DC1 training set. The `EAZY` architecture cannot accept a template set other than the same five basis templates employed by `k-correct` when constructing the DC1 catalogue, but, for consistency with the experimental scope of perfect prior information, `EAZY`'s flexible `all-templates` mode was used to fit the photometric data with a linear combination of the five basis templates. Though `EAZY` can account for uncertainty in the template set by adding in quadrature to the flux errors an empirically derived template error as a function of redshift, we set the template error to zero since the same templates were in fact used to produce the DC1 photometry.

3.1.3 *LePhare*

Photometric Analysis for Redshift Estimate (`LePhare`, Arnouts et al. 1999; Ilbert et al. 2006) uses the χ^2 of Equation 1 to match observed colours with those predicted from a template set. The template set can be semi-empirical or entirely synthetic. The reported photo-z PDF is an arbitrary normalization of the likelihood evaluated on the output redshift grid.

Here we use `LePhare`-v 2.2 with the DC1 template set of Section 2.3.2. Unlike both `BPZ` and `EAZY`, `LePhare` uses generic, SED-independent priors that are not tuned to the DC1 data set.

3.2 Machine Learning-based Approaches

We compared nine data-driven photo-z estimation approaches, eight of which are described in this section and one of which is discussed in Section 3.3. Because the algorithms differ more from one another and the techniques are relative newcomers to the astronomical literature, we provide somewhat more detail about the implementations below.

3.2.1 *ANNz2*

`ANNz2`(Sadeh et al. 2016) supports several machine learning algorithms, including artificial neural networks (ANN), boosted decision tree, and k-nearest neighbour (KNN) regression. In addition to accounting for errors on the input photometry, `ANNz2` uses the KNN-uncertainty estimate of Oyaizu et al. (2008) to quantify uncertainty in the choice of method over multiple runs. Using the Toolkit for Multivariate Data Analysis with ROOT[5], `ANNz2` can return the results of running a single machine learning algorithm, a "best" choice of the results from simultaneously running multiple algorithms (based on evaluation the cumulative distribution functions of validation set objects), or a combination of the results of multiple algorithms weighted by their method uncertainties averaged over multiple runs.

In this study, we used `ANNz2`-v.2.0.4 to output only the result of the ANN algorithm. Photo-z PDFs were produced by running an ensemble of 5 ANNs with a $6:12:12:1$ architecture corresponding to the 6 *ugrizy* inputs, 2 hidden layers with 12 nodes each, and 1 output of redshift. Each of

[5] `http://tmva.sourceforge.net/`

the five ANNs was trained with different random seeds for the initialization of input parameters, reserving half of the training set for validation to prevent overfitting. Undetected galaxies were excluded from the training set, and per-band non-detections in the test set were replaced with the mean magnitude in that band within the entire test set.

3.2.2 *Colour-Matched Nearest-Neighbours*

The colour-matched nearest-neighbours photometric redshift estimator (`CMNN`, Graham et al. 2018) uses a training set of galaxies with known redshifts that has equivalent or better photometry than the test set in terms of quality and filter coverage. For each galaxy in the test set, `CMNN` identifies a colour-matched subset of training galaxies using a threshold in the Mahalanobis distance $D_M = \sum_j^{N_{\rm colours}} (c_j^{\rm train} - c_j^{\rm test})^2/\delta c^2_{\rm test}$ in the space of available colours c, with colour measurement errors $\delta c_{\rm test}$ and $N_{\rm colours} = 5$ colours j defined by the *ugrizy* filters, which defines the set of colour-matched neighbours based on a value of the percent point function (PPF). As an example, for $N_{\rm filt} = 5$ with PPF$= 0.95$, 95 per cent of all training galaxies consistent with the test galaxy will have $D_M < 11.07$. Undetected bands are dropped, thereby reducing the effective $N_{\rm filt}$ for that galaxy. The photo-z PDF of a given test set galaxy is the normalized distribution of redshifts of its colour-matched subset of training set galaxies.

Here, we make two modifications to the implementation of Graham et al. (2018) to comply with the controlled experimental conditions. First, we do not impose non-detections on galaxies fainter than the expected LSST 10-year limiting magnitude nor galaxies bright enough to saturate with LSST's CCDs, instead using all of the photometry for the DC1 test and training sets. Second, we apply the initial colour cut to the training set before calculating the Mahalanobis distance in order to accelerate processing and use a magnitude pseudo-prior as in Graham et al. (2018), but for both we use cut-off values corresponding to the DC1 training set galaxies' colours and magnitudes.

We make an additional adaptation to enable the `CMNN` algorithm to yield accurate photo-z PDFs for all galaxies, as the original Graham et al. (2018) algorithm is optimized for photo-z point estimates and is susceptible to less accurate photo-z PDFs for bright galaxies or those with few matches in colour-space. We use PPF$= 0.95$ rather than PPF$= 0.68$ to generate the subset of colour-matched training galaxies, whose redshifts are weighted by their inverse Mahalanobis distances when composing the photo-z PDF rather than weighting all colour-matched training galaxies equally. Additionally, when the number of colour-matched training set galaxies is less than 20, the nearest 20 neighbours in colour-space are used instead, and the output photo-z PDF is convolved with a Gaussian kernel of variance $\sigma^2_{\rm train}({\rm PPF}_{20}/0.95)^2 - 1$ to account for the corresponding growth of the effective PPF to include 20 neighbors.

3.2.3 *Delight*

`Delight`(Leistedt & Hogg 2017) is a hybrid technique that infers photo-zs with a data-driven model of latent SEDs and a physical model of photometric fluxes as a function of redshift. Generally, machine learning methods rely on representative training data with shared photometric filters, while template-based methods rely on a complete library of templates based on physical models constructed. `Delight` aims to take the best aspects of both approaches by constructing a large collection of latent SED templates (or physical flux-redshift models) from training data, with a template SED library as a guide to the learning of the model, thereby circumventing the machine learning prerequisite of representative training data in the same photometric bands and the template fitting requirement of detailed galaxy SED models. It models noisy observed flux $\hat{\mathbf{F}} = \mathbf{F} + F_b$ as a sum of a noiseless flux plus a Gaussian processes $F_b \sim \mathcal{GP}\left(\mu^F, k^F\right)$ with zero mean function μ^F and a physically motivated kernel k^F that induces realistic correlations in flux-redshift space.

From a template-fitting perspective, each test set galaxy has a posterior $p(z|\hat{\mathbf{F}}) \approx \sum_i p(\hat{\mathbf{F}}|z, T_i)p(z|T_i)p(T_i)$ of redshift z conditioned on noisy flux $\hat{\mathbf{F}}$, where $p(z|T_i)p(T_i)$ captures prior information about the redshift distributions and abundances of the galaxy templates T_i. As in traditional template fitting, each likelihood $p(\hat{\mathbf{F}}|\mathbf{F})$ relates the noisy flux $\hat{\mathbf{F}}$ with the noiseless flux $\mathbf{F}$ predicted by the model of a linear combination of templates, carefully constructed to account for model uncertainties and different normalization of the same SED, plus the Gaussian process term.

The machine learning approach appears in the inclusion of a pairwise comparison term $p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)$ for the prediction of model flux $\mathbf{F}$ at a model redshift z with respect to training set galaxy j with redshift z_j and observed flux $\hat{\mathbf{F}}_j$. Thus the photo-z posterior $p(\hat{\mathbf{F}}|z, T_i) = \int p(\hat{\mathbf{F}}|\mathbf{F})p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)d\mathbf{F}$ may be interpreted as the probability that the training and the target galaxies have the same SED at different redshifts. The flux prediction $p(\mathbf{F}|z, z_j, \hat{\mathbf{F}}_j)$ of the training galaxy at redshift z is modeled via the Gaussian process described above; more detail is provided in Leistedt & Hogg (2017).

In this study, the default settings of `Delight` were used, with the exception that the PDF bins were set to be linearly spaced rather than logarithmically. The Gaussian process was trained using the full DC1 training set. We used the full DC1 template set with a flat prior in magnitude and SED type. Photometric uncertainties from the inputs are propagated into the code, while non-detections for each band are set to the mean of the respective bands.

3.2.4 *FlexZBoost*

`FlexZBoost`(Izbicki & Lee 2017) is built on `FlexCode`, a general-purpose methodology for converting any conditional mean point estimator of z to a conditional density estimator $p(z|\mathbf{x}) \equiv f(z|\mathbf{x})$, where $\mathbf{x}$ here represents our photometric covariates and errors. `FlexZBoost` expands the unknown function $f(z|\mathbf{x}) = \sum_i \beta_i(\mathbf{x})\phi_i(z)$ using an orthonormal basis $\{\phi_i(z)\}_i$. By the orthogonality property, the expansion coefficients $\beta_i(\mathbf{x}) = \mathbb{E}\left[\phi_i(z)|\mathbf{x}\right] \equiv \int f(z|\mathbf{x})\phi_i(z)dz$ are thus conditional means. The expectation value $\mathbb{E}\left[\phi_i(z)|\mathbf{x}\right]$ of the expansion coefficients conditioned on the data is equivalent to the regression of the space of possible redshifts on the space of possible photometry. Thus the expansion coefficients $\beta_i(\mathbf{x})$ can be estimated from the data via regression to yield the conditional density estimate $\widehat{f}(z|\mathbf{x})$.

In this paper, we used `xgboost` (Chen & Guestrin 2016) for the regression; it should, however, be noted that `FlexCode-RF`[6], based on Random Forests, generally performs better on smaller datasets. As our basis $\phi_i(z)$, we choose a standard Fourier basis. The two tuning parameters in our photo-z PDF estimate are the number I of terms in the series expansion and an exponent α that we use to sharpen the computed density estimates $\widetilde{f}(z|\mathbf{x}) \propto \widehat{f}(z|\mathbf{x})^{\alpha}$. Both I and α were chosen in an automated way by minimizing the weighted L_2-loss function (Eq. 5 in Izbicki & Lee 2017) on a validation set comprised of a randomly selected 15 per cent of the DC1 training set. While `FlexCode`'s lossless native encoding stores each photo-z PDF using the basis coefficients $\beta_i(\mathbf{x})$, we discretized the final estimates into 200 linearly spaced redshift bins $0 < z < 2$ to match the consistent output format of the experimental conditions.

3.2.5 GPz

`GPz`(Almosallam et al. 2016a,b) is a sparse Gaussian process-based code, a scalable approximation of full Gaussian Processes (Rasmussen & Williams 2006), that produces input-dependent variance estimates corresponding to heteroscedastic noise. The model assumes a Gaussian posterior probability $p(z|\mathbf{x}) = \mathcal{N}\left(z|\mu(\mathbf{x}), \sigma(\mathbf{x})^2\right)$ of the output redshift z given the input photometry $\mathbf{x}$. The mean $\mu(\mathbf{x})$ and the variance $\sigma(\mathbf{x})^2$ are modeled as functions $f(\mathbf{x}) = \sum_{i=1}^{m} w_i \phi_i(\mathbf{x})$ that are linear combinations of m basis functions $\{\phi_i(\mathbf{x})\}_{i=1}^{m}$ with associated weights $\{w_i\}_{i=1}^{m}$. 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). `GPz`'s variance estimate is composed of a model uncertainty term corresponding to sparsity of the training set photometry and a noise uncertainty term encompassing noisy photometric observations, enabling quantification of any need for more representative or more precise training samples. `GPz` may also weight training set samples by importance according to $|z_{\rm spec} - z_{\rm phot}|/(1+z_{\rm spec})$ to minimize the normalized photo-z point estimate error. However, this function may be adapted to photo-z PDFs, pressuring the model to dedicate more resources to test set galaxies that are not well-represented in the training set.

To smooth the long tail in the distribution of magnitude errors, we use the logarithm of the magnitude errors, improving numerical stability and eliminating the need for constraints on the optimization process. Unobserved magnitudes $x_{\rm u} = \mu_{\rm u} + \Sigma_{\rm uo}\Sigma_{\rm oo}^{-1}(x_{\rm o} - \mu_{\rm o})$ were imputed from observed magnitudes $x_{\rm o}$ and the training set mean μ and covariance Σ using a linear model. This is the optimal expected value of the unobserved variables given the observed ones under the assumption that the distribution is jointly Gaussian; note that this reduces to a simple average if the covariates are independent with $\Sigma_{\rm uo} = 0$. We reserved for validation 20 per cent of the training set and used the Variable Covariance option in `GPz` with 200 basis functions (see Almosallam et al. (2016b) for details), and did not apply cost-sensitive learning options.

[6] `https://github.com/tpospisi/flexcode`; `https://github.com/rizbicki/FlexCoDE`

3.2.6 METAPhoR

Machine-learning Estimation Tool for Accurate Photometric Redshifts (`METAPhoR`, Cavuoti et al. 2017) is based on the Multi Layer Perceptron with Quasi Newton Algorithm (MLPQNA) with the least square error model and Tikhonov L_2-norm regularization (Hofmann & Mathé 2018). Photo-z PDFs are generated by running N trainings on the same training set, or M trainings on M different random samplings of the training set. Upon regression of the test set, the photometry m_{ij} of each test set galaxy j in filter i is perturbed according to $m'_{ij} = m_{ij} + \alpha_i F_{ij} \epsilon$ in terms of the standard normal random variable $\epsilon \sim \mathcal{N}(0, 1)$, a multiplicative constant α_i permitting accommodation of multi-survey photometry, and a bimodal function F_{ij} composed of a polynomial fit of the mean magnitude errors on the binned bands plus a constant term representing the threshold below which the polynomial's noise contribution is negligible (Brescia et al. 2018).

In this work, we used a hierarchical KNN to replace non-detections with values based on their neighbors. The usual cross-validation of redshift estimates and PDFs was also omitted for this study.

3.2.7 SkyNet

`SkyNet`(Graff et al. 2014) employs a neural network based on a second-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 corresponding to the 6 photometric magnitudes and their measurement errors.

`SkyNet`'s classifier mode uses a cross-entropy error function with a 20:40:40 node (all rectified linear units) architecture for each hidden layer and an output layer of 200 nodes corresponding to 200 bins for the PDF, with a softmax activation function to enforce the normalization condition that the probabilities sum to unity. While previous implementations of the code (see Appendix C.3 of Sánchez et al. 2014; Bonnett 2015) implement a sliding bin smoothing, no such procedure was used in this study.

We pre-whitened the data by pegging the magnitudes to (45,45,40,35,42,42) and errors to (20,20,10,5,15,15) for *ugrizy* filters, respectively. To avoid over-fitting, 30 per cent of the training set was reserved for validation, and training was halted as soon as the error rate began to increase on the validation set. The weights were randomly initialized based on normal sampling.

3.2.8 TPZ

Trees for Photo-z (`TPZ`, Carrasco Kind & Brunner 2013; Carrasco Kind & Brunner 2014) uses prediction trees and random forest techniques to estimate photo-z PDFs. `TPZ` recursively splits the training set into branch pairs based on maximizing information gain among a random subsample of features, to minimize correlation between the trees, terminating only when a newly created leaf meets a criterion, such as a leaf size minimum or a variance threshold. The regions in each terminal leaf node correspond to a subsample of the training set with similar properties. Bootstrap samples from the training set photometry and errors are used to build a set of prediction trees.

To run `TPZ`, we replaced non-detections with an approximation of the 1σ detection threshold based on the signal-to-noise-based error forecast of the 10-year LSST data, i. e. $dm = 2.5\log(1 + N/S)$ where $dm \sim 0.7526$ magnitudes for $N/S = 1$ (where N and S are the noise and signal). We calibrated `TPZ` with the Out-of-Bag cross-validation technique (Breiman et al. 1984; Carrasco Kind & Brunner 2013) to evaluate its predictive validity and determine the relative importance of the different input attributes. We grew 100 trees to a minimum leaf size of 5 using the *ugri* magnitudes, all $u-g, g-r, r-i, i-z, z-y$ colours, and the associated errors, as the z and y magnitudes did not show significant correlation with the redshift in our cross-validation. We partitioned our redshift space into 200 bins and smoothed each individual PDF with a smoothing scale of twice the bin size.

3.3 trainZ: a pathological photo-z PDF estimator

We also consider a pathological photo-z PDF estimation method, dubbed `trainZ`, which assigns each test set galaxy a photo-z PDF equal to the normalized redshift distribution $N(z)$ of the training set, according to

$$p(z|\{z_j\}) \equiv \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} \begin{cases} 1 & \text{if } z_k \leq z_i < z_{k+1} \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

Unlike the other methods, the `trainZ` estimator is *independent of the photometric data*, effectively performing a KNN procedure with $k = N_{\rm train}$.

Though `trainZ` is strongly vulnerable to a nonrepresentative training set, it should optimize performance metrics probing the ensemble properties of the galaxy sample, modulo Poisson error due to small sample size, as the training set and test set are drawn from the same underlying population. We will demonstrate its performance under the metrics of Section 4 and discuss it as an illustrative experimental control case in Section 6.1 to highlight the limitations of our evaluation criteria for photo-z PDFs.

4 ANALYSIS

The goal of this study is to evaluate the degree to which photo-z PDFs of each method can be trusted for a generic analysis. The overloaded "$p(z)$" is a widespread abuse of notation that obfuscates this goal, so we dedicate attention to dismantling it here.

Galaxies have redshifts z and photometric data d drawn from a joint probability space $p(z, d)$ in nature, and each observed galaxy i has a *true posterior photo-z PDF* $p(z|d_i)$. There are a number of metrics that can be used to test the accuracy of a photo-z posterior as an estimator of a true photo-z posterior if the true photo-z PDF is known. However, the true photo-z PDF of observed data is not accessible, and existing mock catalogues produce redshift-photometry pairs (z, d) by a deterministic algorithm that does not correspond to a joint probability density from which one can take samples. In these cases there is no "true PDF" for an individual object, and most measures of PDF fidelity will necessarily be restricted to probing the quality of the ensemble of photo-z PDFs. (See §,6.2 for a discussion of how one might circumvent this limitation.)

Before describing the metrics appropriate to the DC1 data set, we outline the philosophy behind our choices. A photo-z PDF estimator derived by method H must be understood as a posterior probability distribution

$$\hat{p}_i^H(z) \equiv p(z|d_i, I_D, I_H), \quad (3)$$

conditioned not only on the photometric data d_i for that galaxy but also on parameters encompassing prior information I_D shared, in our experiment, among all photo-z PDF codes and I_H that will differ depending on the method H used to produce it. To be concrete, I_D takes the form of a training set for the machine learning codes and a template library for the model fitting codes.

The interpretation of the information I_H is more subtle. This investigation is built upon the knowledge that two codes taking the same approach, among choices of model fitting or machine learning, are nonetheless expected to yield different results even if they take the same external prior information I_D. I_H represents the projection of the code's architecture onto the estimated posteriors over redshift, specific to each code, and even the tunable parameters or random seeds of a specific run of a code with a random component. We refer to I_H as the *implicit prior*, in contrast with the training set or template library provided to a given code explicitly by the researcher. In simple terms, the implicit prior is the collection of the many different assumptions, coding choices, algorithm selections, and other implementation details that are specific to each code, the ensemble of which results in differing estimates of redshift when combined with the data and prior information in common to all codes.

The presence of the implicit prior in some sense makes a direct comparison of photo-z PDFs produced by different methods impossible; even if they share the same external 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 effect of differences in prior information I_H specific to each method by using a single training set $I_D^{\rm ML}$ for all machine learning-based codes and a single template library $I_D^{\rm T}$ for all template-based codes. These sets of prior information are carefully constructed to be representative and complete, so we have $I_D \equiv I_D^{\rm ML} \equiv I_D^{\rm T}$ for every method H. Under this assumption, a ratio of posteriors of codes is in effect a ratio of the implicit posteriors $p(z|d_i, I_{H'})$ since the external prior information I_D is present in the numerator and denominator. Thus comparisons of $\hat{p}_i^H(z)$ isolate the effect of the method used to obtain the estimator, which should enable interpretation of the differences between estimated PDFs in terms of the specifics of the method implementations.

The exact implementation of the metrics theoretically depends 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, such as the 200-bin $0 < z < 2$ piecewise constant function used here, the exact bin definitions must affect the result. The piecewise constant format is chosen because of its established presence in the literature, and the choice of 200 bins was motivated by the approximate number of columns expected to be available for storage of photo-z PDFs for the final LSST Project tables.[7] We will discuss the choice of photo-z PDF parameterization further in Section 6.

This analysis is conducted using the qp[8] software package (Malz & Marshall 2018) for manipulating and calculating metrics of univariate PDFs. We present the metrics of photo-z PDFs that address our goals in the sections below. Section 4.1 outlines aggregate metrics of a catalogue of photo-z PDFs, and Section 4.2 presents a metric of individual photo-z PDFs in the absence of true photo-z PDFs. Those seeking a connection to previous comparison studies will find metrics of redshift point estimate reductions of photo-z PDFs in Appendix B and metrics of a science-specific summary statistic heuristically derived from photo-z PDFs in Appendix A.

4.1 Metrics of photo-z PDF ensembles

Because LSST's photo-z PDFs will be used for many scientific applications, some of which require each individual catalogue entry to be accurate, we consider several metrics that probe the population-level performance of the photo-z PDFs. As we have the true redshifts but not true photo-z PDFs for comparison, we remind the reader of the Cumulative Distribution Function (CDF)

$$\mathrm{CDF}[f, q] \equiv \int_{-\infty}^{q} f(z)dz, \tag{4}$$

of a generic univariate PDF $f(z)$, which is used as the basis for several of our metrics. We describe metrics based on the CDF in Section 4.1.1 and metrics of summary statistics thereof in Section 4.1.2.

4.1.1 CDF-based metrics

A quantile of a distribution is the value q at which the CDF of the distribution is equal to Q; percentiles and quartiles are familiar examples of linearly spaced sets of 100 and 4 quantiles, respectively. The quantile-quantile (QQ) plot serves as a graphical visualization for comparing two distributions, where the quantiles of one distribution are plotted against the quantiles of the other distribution, providing an intuitive way to qualitatively assess the consistency between an estimated distribution and a true distribution. The closer the QQ plot is to diagonal, the closer the match between the distributions.

The probability integral transform (PIT)

$$\mathrm{PIT} \equiv \mathrm{CDF}[\hat{p}, z_{true}] \tag{5}$$

is the CDF of a photo-z PDF evaluated at its true redshift, and the distribution of PIT values probes the average accuracy of the photo-z PDFs of an ensemble of galaxies. The distribution of PIT values is effectively the derivative of the QQ plot. A catalogue of accurate photo-z PDFs should have a PIT distribution that is uniform $U(0, 1)$, and deviations from flatness are interpretable: overly broad photo-z PDFs induce underrepresentation of the lowest and highest PIT values, whereas overly narrow photo-z PDFs induce overrepresentation of the lowest and highest PIT values. Catastrophic outliers with a true redshift outside the support of its photo-z PDF have PIT ≈ 0 or PIT ≈ 1.

The PIT distribution has been used to quantify the performance of photo-z PDF methods in the past (e. g. Bordoloi et al. 2010; Polsterer et al. 2016; Tanaka et al. 2018). Tanaka et al. (2018) use the histogram of PIT values as a diagnostic indicator of overall code performance, while Freeman et al. (2017) independently define the PIT and demonstrate how its individual values may be used both to perform hypothesis testing (via, e. g. the KS, CvM, and AD tests; see below) and to construct QQ plots. Following Kodra & Newman (in prep.) we define the PIT-based catastrophic outlier rate as the fraction of galaxies with PIT < 0.0001 or PIT > 0.9999, which should total 0.0002 for an ideal uniform distribution.

4.1.2 Summary statistics of CDF-based metrics

We evaluate a number of quantitative metrics derived from the visually interpretable QQ plot and PIT histogram, built on the Kolmogorov-Smirnov (KS) statistic

$$\mathrm{KS} \equiv \max_{z} \left(\left| \mathrm{CDF}[\hat{f}, z] - \mathrm{CDF}[\tilde{f}, z] \right| \right), \tag{6}$$

interpretable as the maximum difference between the CDFs of the empirical distribution of PIT values for the test sample $\hat{f}(z)$ and a reference distribution $\tilde{f}(z)$, in this case $U(0, 1)$, for the ideal distribution of PIT values. We also consider two variants of the KS statistic. A cousin of the KS statistic, the Cramer-von Mises (CvM) statistic

$$\mathrm{CvM}^2 \equiv \int_{-\infty}^{+\infty} \left(\mathrm{CDF}[\hat{f}, z] - \mathrm{CDF}[\tilde{f}, z] \right)^2 \mathrm{dCDF}[\tilde{f}, z] \tag{7}$$

is the mean-squared difference between the CDFs of an approximate and true PDF. The Anderson-Darling (AD) statistic

$$\mathrm{AD}^2 \equiv N_{tot} \int_{-\infty}^{+\infty} \frac{\left(\mathrm{CDF}[\hat{f}, z] - \mathrm{CDF}[\tilde{f}, z] \right)^2}{\mathrm{CDF}[\tilde{f}, z](1 - \mathrm{CDF}[\tilde{f}, z]} \mathrm{dCDF}[\tilde{f}, z] \tag{8}$$

is a weighted mean-squared difference featuring enhanced sensitivity to discrepancies in the tails of the distribution. In anticipation of a substantial fraction of galaxies having PIT of 0 or 1, a consequence of catastrophic outliers, we evaluate the AD statistic with modified bounds of integration $(0.01, 0.99)$ to exclude those extremes in the name of numerical stability.

4.2 Conditional Density Estimate (CDE) Loss: a metric of individual photo-z PDFs

The Buzzard simulation process precludes testing the degree to which samples from our photo-z posteriors reconstruct the space of $p(z, \mathrm{data})$. To the knowledge of the authors, there is only one metric that can be used to evaluate the performance of individual photo-z PDF estimators in the absence of true photo-z posteriors. The conditional density

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

estimation (CDE) loss is an analogue to the familiar root-mean-square-error used in conventional regression, defined as

$$L(f, \widehat{f}) \equiv \int \int (f(z|\mathbf{x}) - \widehat{f}(z|\mathbf{x}))^2 \mathrm{d}z \mathrm{d}P(\mathbf{x}), \tag{9}$$

where $f(z|\mathbf{x})$ is the true photo-z PDF that we do not have and $\widehat{f}(z|\mathbf{x})$ is an estimate thereof, in terms of the photometry $\mathbf{x}$. (See Section 3.2.4 for a review of the notation.) We estimate the CDE loss via

$$\hat{L}(f, \widehat{f}) = \mathbb{E}_{\mathbf{X}} \left[\int \widehat{f}(z \mid \mathbf{X})^2 dz \right] - 2\mathbb{E}_{\mathbf{X},Z} \left[\widehat{f}(Z \mid \mathbf{X}) \right] + K_f, \tag{10}$$

where the first term is the expectation value of the photo-z posterior with respect to the marginal distribution of the photometric covariates $\mathbf{X}$, the second term is the expectation value with respect to the joint distribution of $\mathbf{X}$ and the space Z of all possible redshifts, and the third term K_f is a constant depending only upon the true conditional densities $f(z|\mathbf{x})$. We may estimate these expectations empirically on the test or validation data (Eq. 7 in Izbicki et al. 2017) without knowledge of the true densities.

5 RESULTS

We begin with a demonstrative visual inspection of the photo-z PDFs produced by each code for individual galaxies. Figure 1 shows the photo-z PDFs for four galaxies chosen as examples of photo-z PDF archetypes: a narrow unimodal PDF, a broad unimodal PDF, a bimodal PDF, and a multimodal PDF. We reiterate that under our idealized experimental conditions, differences between codes are the isolated signature of the implicit prior due to the method by which the photo-z PDFs were derived.

The most striking differences between codes are the small-scale features induced by the interaction between the shared piecewise constant parameterization of 200 bins $0 < z < 2$ of Section 4 and the smoothing conditions or lack thereof in each algorithm. The $\mathrm{d}z = 0.01$ redshift resolution is sufficient to capture the broad peaks of faint galaxies' photo-z PDFs with large photometric errors but is too broad to resolve the narrow peaks for bright galaxies' photo-z PDFs with small photometric errors. This observation is consistent with the findings of Malz et al. (2018) that the piecewise constant parameterization underperforms in the presence of small-scale structures.

However, the shared small-scale features of `ANNz2`, `METAPhoR`, `CMNN`, and `SkyNet` are a result of various weighted sums of the limited number of training set galaxies with colours similar to those of the test set galaxy in question, with behavior closer to classification than regression in the case of `ANNz2`. The settings used on `GPz` in this work forced broadening of the single Gaussian to cover the multimodal redshift solutions of the other codes.

5.1 Performance on photo-z PDF ensembles

The histogram of PIT values, QQ plot, and QQ difference plot relative to the ideal diagonal are provided in Figure 2, showcasing the biases and trends in the average accuracy of the photo-z PDFs for each code. The high QQ values (i. e. more high than low PIT values) of `BPZ`, `CMNN`, `Delight`, `EAZY`, and `GPz` indicate photo-z PDFs biased low, and the low QQ values (more low than high PIT values) of `SkyNet` and `TPZ` indicate photo-z PDFs biased high. The gray shaded band marks the 2σ variance in PIT values found using the `trainZ` algorithm with a bootstrap resampling of the training set and a sample size of 30,000 galaxies, representing a very conservative estimate of the representative training sample size, and thus an approximate minimal error significance compared to ideal performance. Deviations in the PIT histograms outside of this range show that significant biases are present for some codes.

Table 2. The catastrophic outlier rate as defined by extreme 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 PIT$< 10^{-4}$ or $>$0.9999
`ANNz2`	0.0265
`BPZ`	0.0192
`Delight`	0.0006
`EAZY`	0.0154
`FlexZBoost`	0.0202
`GPz`	0.0058
`LePhare`	0.0486
`METAPhoR`	0.0229
`CMNN`	0.0034
`SkyNet`	0.0001
`TPZ`	0.0130
`trainZ`	0.0002

The PIT histograms of `Delight`, `CMNN`, `SkyNet`, and `TPZ` feature an underrepresentation of extreme values, indicative of overly broad photo-z PDFs, while the overrepresentation of extreme values for `METAPhoR` indicate overly narrow photo-z PDFs. These five codes in particular have a free parameter for bandwidth, which may be responsible for this vulnerability, in spite of the opportunity for fine-tuning with perfect prior information. `FlexZBoost`'s "sharpening" parameter (described in Section 3.2.4) played a key role in diagonalizing the QQ plot, indicating a common avenue for improvement in the approaches that share this type of parameter. On the other hand, the three purely template-based codes, `BPZ`, `EAZY`, and `LePhare`, do not exhibit much systematic broadening or narrowing, which may indicate that complete template coverage effectively defends from these effects.

Close inspection of the extremes at PIT values of 0 and 1 reveal spikes in the first and last bin of the PIT histogram for some codes in Figure 2, corresponding to catastrophic outliers where the true redshift lies outside of the support of the $p(z)$. The catastrophic outlier rates are provided in Table 2. As expected, `trainZ` achieves precisely the 0.0002 value expected of an ideal PIT distribution. `ANNz2`, `FlexZBoost`, `LePhare`, and `METAPhoR` have notably high catastrophic outlier rates > 0.02, exceeding 100 times the ideal PIT rate, meriting further investigation.

Figure 3 highlights the relative values of the KS, CvM, and AD test statistics calculated by comparing the PIT distribution and a uniform distribution $U(0, 1)$. `METAPhoR` and `LePhare` perform well under the AD but poorly under the

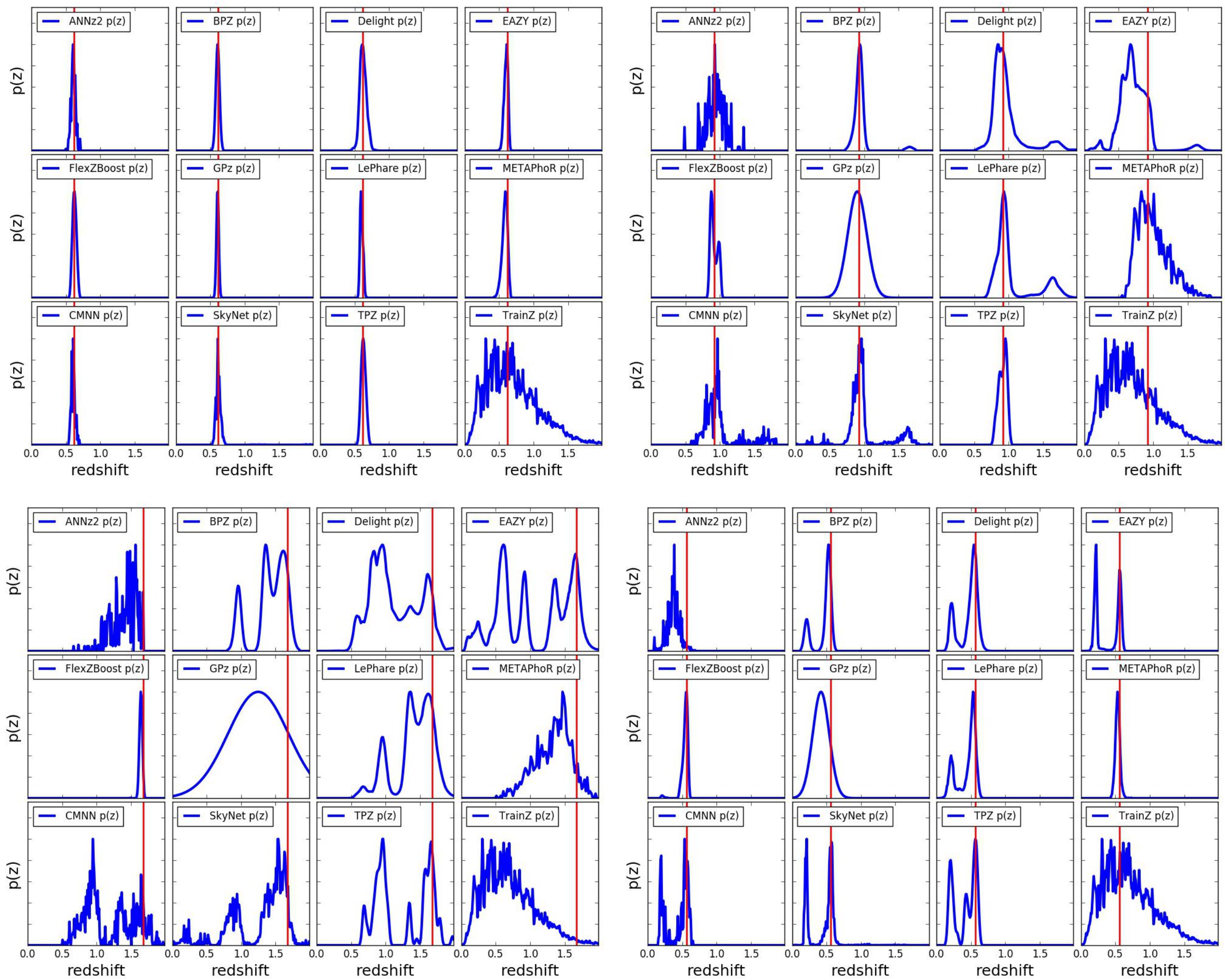

Figure 1. The individual photo-z PDFs (blue) distributions produced by the twelve codes (small panels) on four exemplary galaxies' photometry (large panels) with different true redshifts (red). The photo-z PDFs of all codes share some features for the example galaxies due to physical colour degeneracies and photometric errors: tight unimodal $p(z)$ (upper left), broad unimodal $p(z)$ (upper right), bimodal $p(z)$ (lower right), and complex/multimodal $p(z)$ (lower left). The diverse algorithms and implementations induce differences in small-scale structure and sensitivity to physical systematics.

KS and CvM due to their high catastrophic outlier rates. `ANNz2` and `FlexZBoost` are the top scorers under these metrics of the PIT distribution. `ANNz2`'s strong performance can be attributed to an aspect of the training process in which training set galaxies with a PIT that more closely matches the percentiles of the DC1 training set's redshift distribution are upweighted; in effect, these quantile-based metrics were part of the algorithm itself that may or may not serve it well under more realistic experimental conditions. Similar to what was done for the PIT histograms in Figure 2, we create bootstrap training samples of 30,000 galaxies for use with `trainZ` in order to estimate a conservative statistical floor that we would expect in real data. No code reaches this idealized floor, indicating that all codes suffer some degradation from the ideal when employing their implicit priors, though `ANNz2`, `FlexZBoost`, and `GPz` are within a factor of two.

5.2 Performance on individual photo-z PDFs

The values of the CDE loss statistic of individual photo-z PDF accuracy are provided in Table 3. It is worth noting that strong performance on the CDE loss, corresponding to lower values of the metric, should imply strong performance on the other metrics, though the inverse is not necessarily true. Thus the CDE loss is the most effective metric for generic science cases.

Of the metrics we were able to consider in this experiment, **the CDE Loss is the only metric that can appropriately penalize the pathological `trainZ`**. Additionally, it favors `CMNN` and `FlexZBoost`, the latter of which is optimized for this metric.

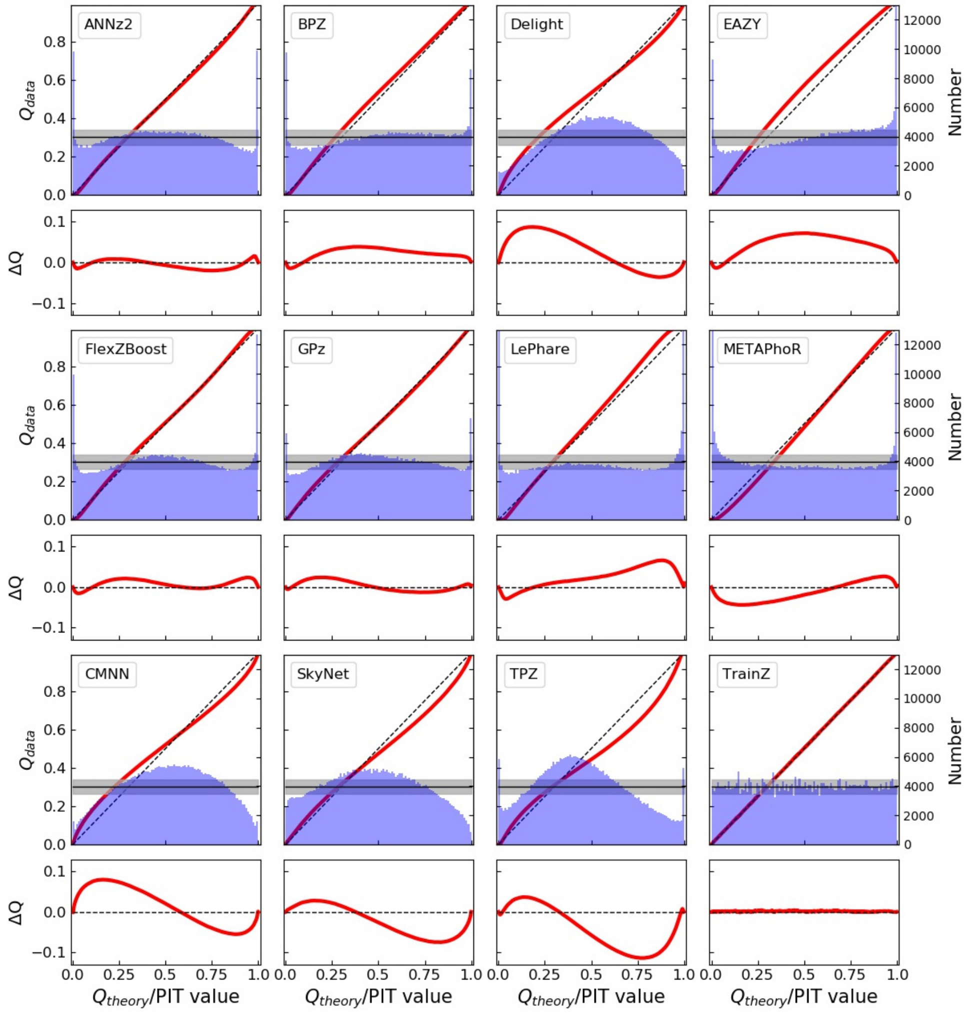

Figure 2. The QQ plot (red) and PIT histogram (blue) of the photo-z PDF codes (panels) along with the ideal QQ (black dashed diagonal) and ideal PIT (gray horizontal) curves, as well as a difference plot for the QQ difference from the ideal diagonal (lower inset). The gray shaded region indicates the 2σ range from a bootstrap resampling of the training set with a size of 30,000 galaxies using `trainZ`. The twelve codes exhibit varying degrees of four deviations from perfection: an overabundance of PIT values at the centre of the distribution indicate a catalogue of overly broad photo-z PDFs, an excess of PIT values at the extrema indicates a catalogue of overly narrow photo-z PDFs, catastrophic outliers manifest as overabundances at PIT values of 0 and 1, and asymmetry indicates systematic bias, a form of model misspecification. Values in excess of the 2σ shaded region show that for some codes these errors will be significant given expected training sample sizes.

6 DISCUSSION AND FUTURE WORK

In contrast with other photo-z PDF comparison papers that have aimed to identify the "best" code for a given survey, we have focused on the somewhat more philosophical questions of how to assess photo-z PDF methods and how to interpret differences between codes in terms of photo-z PDF performance. In Section 6.1, we reframe the strong performance of our pathological photo-z PDF technique, `trainZ`, as a cautionary tale about the importance of choosing appropriate comparison metrics. In Section 6.2, we outline the experiments we intend to build upon this study In Section 6.3, we discuss the enhancements of the mock data set that will be necessary to enable the future experiments.

6.1 Interpretation of metrics

We remind the reader that contributed codes were given a goal of obtaining accurate photo-z PDFs, not an accurate stacked estimator of the redshift distribution, so we do not expect the same codes to necessarily perform well for both classes of metrics. Indeed, the codes were optimized for their interpretation of our request for "accurate photo-z PDFs," and we expect that the implementations would have been adjusted had we requested optimization of the traditional metrics of Appendices A and B.

Furthermore, our metrics are not necessarily able to assess the fidelity of individual photo-z PDFs relative to true posteriors: in the absence of a "true PDF" from which redshifts are drawn, it is difficult to construct metrics to measure performance for individual galaxies rather than ensem-

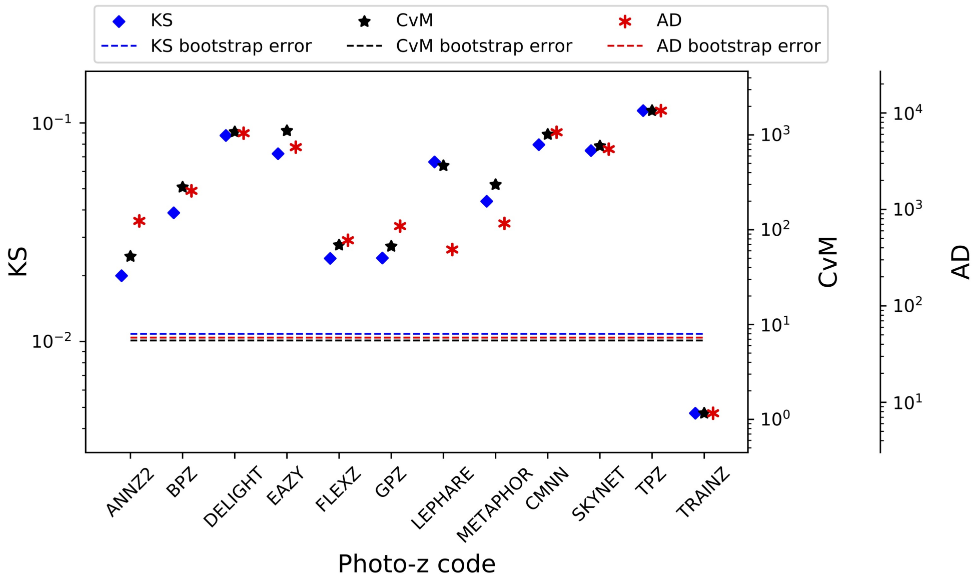

Figure 3. A visualization of the Kolmogorov-Smirnoff (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the PIT distributions. There is generally good agreement between these statistics, with differences corresponding to the codes with outstanding catastrophic outlier rates, a reflection in the differences in how each statistic weights the tails of the distribution. Horizontal lines indicate the level of uncertainty found by bootstrapping a training set sample of 30,000 galaxies using `trainZ`; none of the codes reach this conservative ideal floor in expected uncertainty.

Table 3. CDE loss statistic of the individual photo-z PDFs for each code. A lower value of the CDE loss indicates more accurate individual photo-z PDFs, with `CMNN` and `FlexZBoost` performing best under this metric.

Photo-z Code	CDE Loss
`ANNz2`	-6.88
`BPZ`	-7.82
`Delight`	-8.33
`EAZY`	-7.07
`FlexZBoost`	-10.60
`GPz`	-9.93
`LePhare`	-1.66
`METAPhoR`	-6.28
`CMNN`	-10.43
`SkyNet`	-7.89
`TPZ`	-9.55
`trainZ`	-0.83

bles. (The CDE Loss metric of section 4.2 is an exception to this rule.) A lack of appropriate metrics more sophisticated than the CDE Loss remains an open issue for science cases requiring accurate individual galaxy PDFs. The metric-specific performance demonstrated in this paper implies that we may need multiple photo-z PDF approaches tuned to each metric in order to maximize returns over all science cases in large upcoming surveys.

The `trainZ` estimator of Section 3.3, which assigns every galaxy a photo-z PDF equal to $N(z)$ of the training set, is introduced as an experimental control or null test to demonstrate this point via *reductio ad absurdum*. Because our training set is perfectly representative of the test set, $N(z)$ should be identical for both sets down to statistical noise. **We make the alarming observation that `trainZ`, the experimental control, outperforms all codes on the CDF-based metrics, and all but one code on the $N(z)$ based statistics.** The PIT and other CDF-based metrics upon which modern photo-z PDF comparisons are built (Bordoloi et al. 2010; Polsterer et al. 2016; Tanaka et al. 2018) can be gamed by a trivial estimator that yields only an affirmation of prior knowledge uninformed by the data. In other words, such ensemble metrics are not appropriate for the task of selecting photo-z PDF codes for analysis pipelines.

The CDE loss and point estimate metrics appropriately penalize `trainZ`'s naivete. As shown in Appendix B, `trainZ` has identical $ZPEAK$ and $ZWEIGHT$ values for every galaxy, and thus the photo-z point estimats are constant as a function of true redshift, 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 5.2, distinguishes this metric from those of the photo-z PDF ensemble and stacked estimator of the redshift distribution, despite their prevalence in the photo-z literature.

In summary, context is crucial to defend against deceptively strong performers such as `trainZ`; **the best photo-z PDF method is the one that most effectively achieves our science goals**, not the one that performs best on a metric that does not reflect those goals. In the absence of a single scientific motivation or the information necessary for a principled metric definition, we must consider many metrics and be critical of the information transmitted by each.

6.2 Extensions to the experimental design

The work presented in this paper is only a first step in assessing photo-z PDF approaches and moving toward an improved photometric redshift estimator. Here we discuss the next steps for subsequent investigations.

This initial paper explores photo-z PDF code performance in idealized conditions with perfect catalogue-based photometry and representative training data, but the resilience of each code to such realistic imperfections in prior information has not yet been evaluated. A top priority for a follow-up study is to test realistic forms of incomplete, erroneous, and non-representative template libraries and training sets as well as the impact of other forms of external priors that must be ingested by the codes, major concerns in Newman et al. (2015); Masters et al. (2017). Outright redshift failures due to emission line misidentification or noise spikes may be modeled by the inclusion of a small number of high-confidence yet false redshifts. We plan to perform a full sensitivity analysis on a realistically incomplete training set of spectroscopic galaxies, modeling the performance of spectrographs, emission-line properties, and expected signal-to-noise to determine which potential training set galaxies are most likely to be excluded.

Appendix A only addresses the stacked estimator of the redshift distribution of the entire galaxy catalogue rather than subsets in bins, tomographic or otherwise. The effects of tomographic binning scheme will be explored in a dedicated future paper, including propagation of redshift uncertainties in a set of fiducial tomographic redshift bins in order to estimate impact on cosmological parameter estimation.

Sequels to this study will also address some shortcomings of our experimental procedure. The fixed redshift grid shared between the codes may have unfairly penalized codes with a different native parameterization, as precision is lost when converting between formats. Performance on the (admittedly small) population of sharply peaked photo-z PDFs may have been suppressed across all codes due to the insufficient resolution of the redshift grid. In light of the results of Malz et al. (2018), in future analyses we plan to switch from a fixed grid to the quantile parameterization or to permit each code to use its native storage format under a shared number of parameters.

Section 4 discussed the difficulty in evaluating PDF accuracy for individual objects with known (z, d) information but without a known $p(z, d)$. In a follow-up study, we will generate mock data probabilistically, yielding true PDFs in addition to true redshifts and photometric data. This future data set will enable tests of PDF accuracy for individual galaxies rather than solely ensembles.

6.3 Realistic mock data

To make optimal use of the LSST data for cosmological and other astrophysical analyses of the LSST-DESC Science Roadmap, future investigations that build upon this one will require a more sophisticated set of galaxy photometry and redshifts. This initial paper explored a data set that was constructed at the catalogue level, with no inclusion of the complications that come from measuring photometry from images. Future data challenges will move to catalogues constructed from mock images, including the complications of deblending, sensor inefficiencies, and heterogeneous observing conditions, all anticipated to affect the measured colours of LSST's galaxy sample (Dawson et al. 2016).

The DC1 galaxy SEDs were linear combinations of just five basis SED templates, but a next generation of data for photo-z PDF investigations must include a broader range of physical properties. Though we only considered $z < 2$ here, LSST 10-year data will contain $z > 2$ galaxies, plagued by fainter apparent magnitudes and anomalous colours due to stellar evolution. A subsequent study must also have a data set that includes low-level active galactic nuclei (AGN) features in the SEDs, which perturb colours and other host galaxy properties. An observational degeneracy between the Lyman break of a $z \sim 2 - 3$ galaxy from the Balmer break of a $z \sim 0.2 - 0.3$ galaxy is a known source of catastrophic outliers (Massarotti et al. 2001) that was not effectively included in this study. To gauge the sensitivity of photo-z PDF estimators to catastrophic outliers, our data set must include realistic high-redshift galaxy populations.

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).

7 CONCLUSION

This paper compares twelve photo-z PDF codes under controlled experimental conditions of representative and complete prior information to set a baseline for an upcoming sensitivity analysis. This work isolates the impact on metrics of photo-z PDF accuracy due to the estimation technique as opposed to the complications of realistic physical systematics of the photometry. Though the mock data set of this investigation did not include true photo-z posteriors for comparison, **we interpret deviations from perfect results given perfect prior information as the imprint of the implicit assumptions underlying the estimation approach**.

We evaluate the twelve codes under science-agnostic metrics both established and emerging to stress-test the ensemble properties of photo-z PDF catalogues derived by each method. In appendices, we also present metrics of point estimates and a prevalent summary statistic of photo-z PDF catalogues used in cosmological analyses to enable the reader to relate this work to studies of similar scope. We observe that no one code dominates in all metrics, and that the standard metrics of photo-z PDFs and the stacked estimator of the redshift distribution can be gamed by a very simplistic procedure that asserts the prior over the data. We emphasize to the photo-z community that **metrics used to vet photo-z PDF methods must be scrutinized to ensure they correspond to the quantities that matter to our science**.

Acknowledgments

This paper has undergone internal review in the LSST Dark Energy Science Collaboration, the internal reviewers were: Daniel Gruen, Markus Rau, and Michael Troxel.

Author contributions are listed below.

S.J. Schmidt: Co-led the project. (conceptualization, data curation, formal analysis, investigation, methodology,

project administration, resources, software, supervision, visualization, writing – original draft, writing – review & editing)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

E. Nourbakhsh: Ran and optimized TPZ code and wrote a subsection of Section 2 for TPZ

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

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

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

R. Zhou: Optimized and ran EAZY and contributed to the draft

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

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

personal funding sources SJS acknowledges support from DOE grant DE-SC0009999 and NSF/AURA grant N56981C. AIM acknowledges support from the Max Planck Society and the Alexander von Humboldt Foundation in the framework of the Max Planck-Humboldt Research Award endowed by the Federal Ministry of Education and Research. During the completion of this work, AIM was advised by David W. Hogg and was supported by National Science Foundation grant AST-1517237.

In addition to packages cited in the text, analyses performed in this paper used the following software packages: `Numpy` and `Scipy` (Oliphant 2007), `Matplotlib` (Hunter 2007), `Seaborn` (Waskom et al. 2017), `minFunc` (Schmidt 2005), `qp` (Malz & Marshall 2018; Malz et al. 2018), `pySkyNet` (Bonnett 2016), and `photUtils` from the LSST simulations package (Connolly et al. 2014).

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

APPENDIX A: EVALUATION OF THE REDSHIFT DISTRIBUTION

Perhaps the most popular application of photo-z PDFs is the estimation of the overall redshift distribution $N(z)$, a quantity that enters some cosmological calculations and the true value of which is known for the DC1 data set and will be denoted as $\tilde{N}(z)$. In terms of the prior information provided to each method, the true redshift distribution satisfies the tautology $\tilde{N}(z) = p(z|I_D)$ due to our experimental set-up; because the DC1 training and template sets are representative and complete, I_D represents a prior that is also equal to the truth. In this ideal case of complete and representative prior information, the method that would give the best approximation to $\hat{N}(z)$ 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 $\hat{p}_i(z) = \tilde{N}(z)$ for all i; the inclusion of any information from the photometry would only introduce noise to the optimal result of returning the prior. This is the exact estimator, `trainZ`, that we have described in Section 3.3, and which will serve as an experimental control.

A1 Metrics of the stacked estimator of the redshift distribution

"Stacking" according to

$$\hat{N}^H(z) \equiv \frac{1}{N_{tot}} \sum_{i}^{N_{tot}} \hat{p}_i^H(z) \quad \text{(A1)}$$

is the most widely used method for obtaining $\hat{N}^H(z)$ as an estimator of the redshift distribution from photo-z PDFs derived by a method H. While the stacked estimator of the redshift distribution violates the mathematical definition of statistical independence and is thus not formally correct[9], we use it as a basis for comparison of photo-z PDF methods under the untested assumption that the response of our metrics of $\hat{N}^H(z)$ will be analogous to the same metrics applied to a principled estimator of the redshift distribution.

As $N(z)$ is itself a univariate PDF, we apply the metrics of the previous sections to it as well. We additionaly calculate the first three moments

$$\langle z^m \rangle \equiv \int_{-\infty}^{\infty} z^m N(z) dz \quad \text{(A2)}$$

of the estimated redshift distribution $\hat{N}^H(z)$ for each code and compare them to the moments of the true redshift distribution $\tilde{N}(z)$. Under the assumption that the stacked estimator is unbiased, a superior method minimizes the difference between the true and estimated moments.

A2 Performance on the stacked estimator of the redshift distribution

Figure A1 shows the stacked estimator $\hat{N}(z)$ of the redshift distribution for each code compared to the true redshift distribution $\tilde{N}(z)$, where the stacked estimator has been smoothed for each code in the plot using a kernel density estimate (KDE) with a bandwidth chosen by Scott's Rule (Scott 1992) in order to minimize visual differences in small-scale features; the quantiative statistics, however, are calculated using the empirical CDF which is not smoothed.

Many of the codes, including all the model-fitting approaches and `ANNz2`, `GPz`, `METAPhoR`, and `SkyNet` from the data-driven camp, overestimate the redshift density at $z \sim 1.4$. This behavior is a consequence of the 4000 Å break passing through the gap between the z and y filters, which induces a genuine discontinuity in the $z - y$ colour as a function of redshift that can sway the photo-z PDF estimates in the absence of bluer spectral features.

`ANNz2`, `GPz`, and `METAPhoR` feature exaggerated peaks and troughs relative to the training set, a potential sign of overtraining. Further investigation on overtraining is needed, if present this is an obstacle that may be overcome with adjustment of the implementation.

As expected, `trainZ` perfectly recovers the true redshift distribution: as the training sample is selected from the same underlying distribution as the test set, the redshift distributions are identical, up to Poisson fluctuations due to the finite number of sample galaxies. `CMNN` is also in excellent agreement for similar reasons: with a representative training sample of galaxies spanning the colour-space, the sum of the colour-matched neighbour redshifts should return the true redshift distribution. `FlexZBoost` and `TPZ` also perform superb recovery of the true redshift distribution, with only a slight deviation at $z \sim 1.4$. Our metrics, however, cannot discern whether these four approaches, as well as `Delight`, are spared the $z \sim 1.4$ degeneracy in $\hat{N}(z)$ because they have more effectively used information in the data or if the impact is simply washed out by the stacked estimator's effective average over the test set galaxy sample. See Appendix B for further discussion of the $z \sim 1.4$ issue.

Figure A2 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. The horizontal lines show the the result of a bootstrap resampling of the training set using 30,000 samples for `trainZ`, representing a conservative idealized limit on expected performance for a modest-sized representative training set of galaxies, as mentioned in Section 5.1. The AD bootstrap statistic is elevated due to its sensitivity to the tails of distributions. The stacked estimators of the redshift distribution for `CMNN` and `trainZ` best estimate $\tilde{N}(z)$ under these metrics, whereas `EAZY`, `LePhare`, `METAPhoR`, and `SkyNet` underperform; `BPZ`, `GPz`, and `TPZ` are within a factor of two of the conservative limit for all statistics. It is unsurprising that `CMNN` scores well, as with a nearly complete and representative training set choosing neighbouring points in colour/magnitude space to construct an estimator should lead to excellent agreement in the final $\hat{N}(z)$.

It is, however, surprising that `TPZ` does well on $\hat{N}(z)$ given its poor performance on the ensemble photo-z PDFs, especially knowing that `TPZ` was optimized for photo-z PDF ensemble metrics rather than the stacked estimator of the redshift distribution. A possible explanation is the choice of smoothing parameter chosen during validation, which affects photo-z PDF widths as well as overall redshift bias and could be modified to improve performance under the photo-z PDF metrics.

We calculated the first three moments of the stacked $\hat{N}(z)$ distribution of all galaxies and compared it to the moments of the true redshift distribution. Figure A3 shows the residuals of the moments for all codes. Accuracy of the moments varies widely between codes, raising concerns about the propagation to cosmological analyses. The DESC SRD (The LSST Dark Energy Science Collaboration et al. 2018) lists stringent requirements on how well the mean and variance of tomographic redshift bins must be known for each of the main DESC science cases. We indicate the Year 10 (Y10) requirements assuming our true mean redshift of $z = 0.701$ as dashed lines. In this study with representative training data, `ANNz2`, `CMNN`, `TPZ`, and our pathological `trainZ` estimator meet the Y10 requirement on the mean redshift. Only `ANNz2`, `CMNN`, and `trainZ` meet both requirements. One should be concerned that many codes fail to meet this ambitious limit under perfect prior information because all codes are anticipated to do no better under realistically imperfect prior information, and indicates that additional calibration to remove these systematic offsets (e. g. Newman 2008) will likely be necessary in order to meet these stringent goals.

`SkyNet` exhibits redshift bias in Figure A1 and is a clear outlier in the first moment of $\hat{N}(z)$ in Figure A3. The `SkyNet`

[9] Malz & Hogg (prep) shows how the stacking procedure can lead to bias in the estimate of $N(z)$ and presents a principled alternative to this commonly employed method.

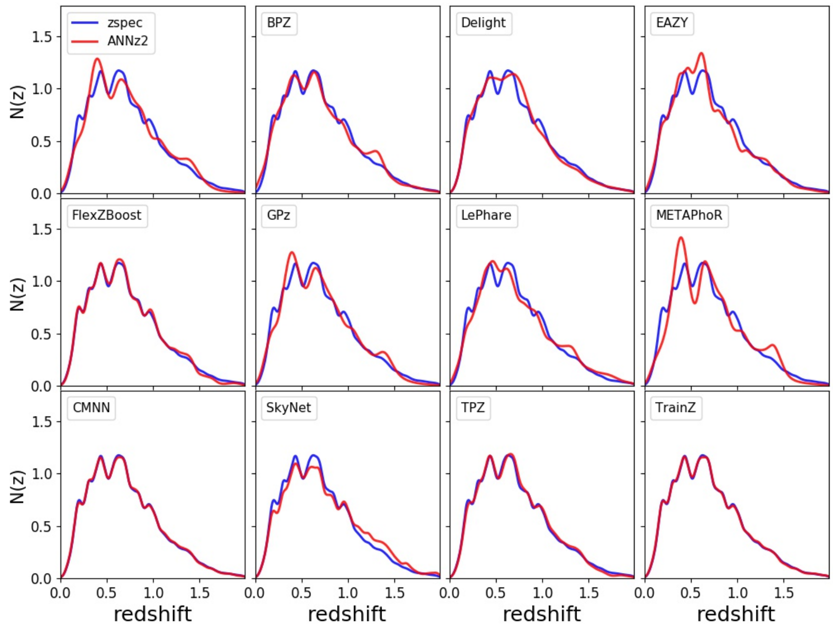

Figure A1. The smoothed stacked estimator $\hat{N}(z)$ of the redshift distribution (red) produced by each code (panels) compared to the true redshift distribution $\tilde{N}(z)$ (blue). Varying levels of agreement are seen among the codes, with the smallest deviations for `CMNN`, `FlexZBoost`, `TPZ`, and `trainZ`.

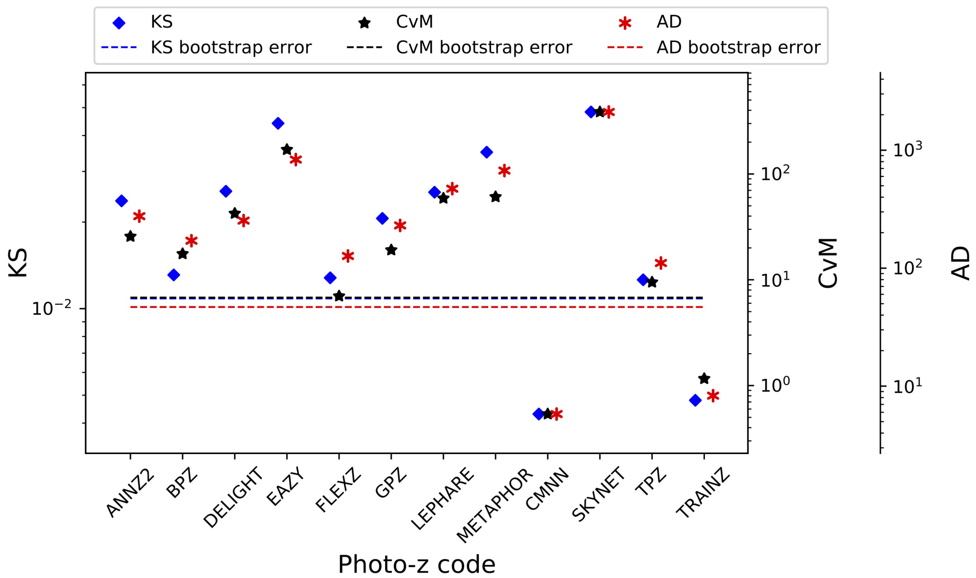

Figure A2. A visualization of the Kolmogorov-Smirnoff (KS, blue diamond), Cramer-von Mises (CvM, black star), and Anderson-Darling (AD, red asterisk) statistics for the $\hat{N}(z)$ distributions. Horizontal lines indicate the statistic values (including uncertainty) achieved using `trainZ` via bootstrap resampling a training set containing 30,000 redshifts. We make the reassuring observation that these related statistics do not disagree significantly with one another. `CMNN` outperforms the control case, `trainZ`, and several codes are within a factor of two of this conservative idealized limit. `SkyNet` scores poorly due to an overall bias in its redshift predictions.

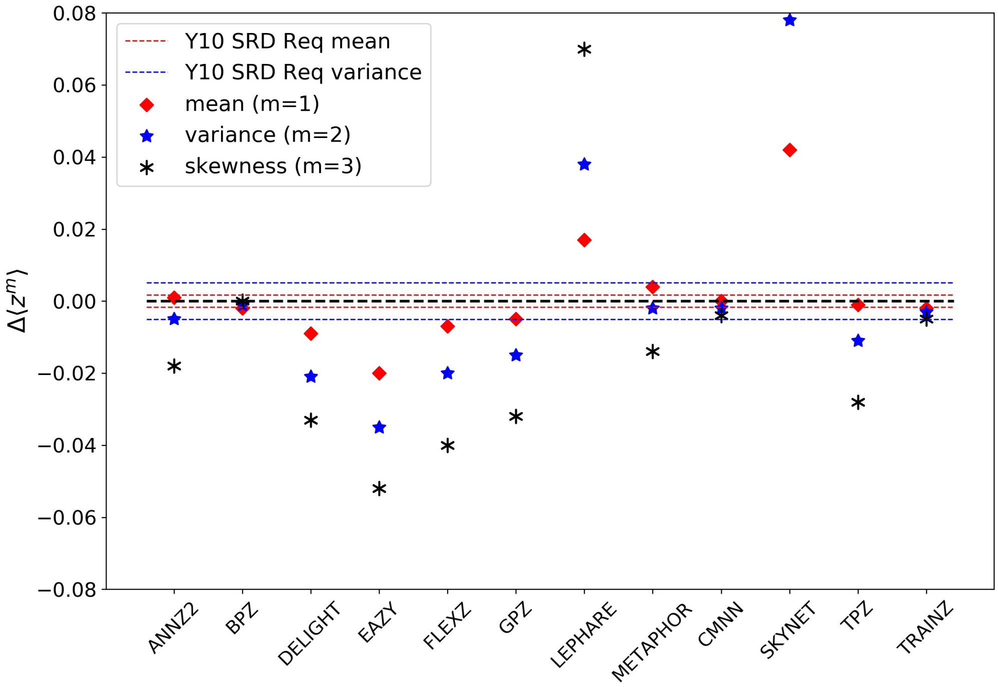

Figure A3. Residuals of the first three moments of the stacked $\hat{N}(z)$ distribution. Red and blue horizontal lines indicate the Year 10 DESC SRD requirements on accuracy of the mean and variance respectively. Only a small number of codes are able to meet these specifications even with perfect training data.

algorithm employs a random subsampling of the training set without testing that the subset is representative of the full population, and the implementation used here does not upweight rarer low- and high-redshift galaxies, as in Bonnett (2015), suggesting a possible cause that may be addressed in future work.

APPENDIX B: Photo-z POINT ESTIMATION AND METRICS

While this work assumes that science applications value the information of the full photo-z PDF, we present conventional metrics of photo-z point estimates as a quick and dirty visual diagnostic tool and to facilitate direct comparisons to historical studies.

B1 Reduction of photo-z PDFs to point estimates

Though we acknowledge that many of the codes can also return a native photo-z point estimate, we put all codes on equal footing by considering two generic photo-z point estimators, the mode z_{PEAK} and main-peak-mean z_{WEIGHT} (Dahlen et al. 2013), a weighted mean within the bounds of the main peak, as identified by the roots of $p(z) - 0.05 \times z_{PEAK}$. Though z_{WEIGHT} neglects information in a secondary peak of e. g. a bimodal distribution, it avoids the pitfall of reducing the photo-z PDF to a redshift between peaks where there is low probability.

B2 Metrics of photo-z point estimates

We calculate the commonly used point estimate metrics of the overall intrinsic scatter, bias, and catastrophic outlier rate, defined in terms of the standard error $e_z \equiv (z_{PEAK} - z_{\rm true})/(1 + z_{\rm true})$. Because the standard deviation of the photo-z residuals is sensitive to outliers, we define the scatter in terms of the Interquartile Range (IQR), the difference between the 75th and 25th percentiles of the distribution of e_z, imposing the scaling $\sigma_{\rm IQR} = {\rm IQR}/1.349$ to ensure that the area within $\sigma_{\rm IQR}$ is the same as that within one standard deviation from a standard Normal distribution. We also resist the effect of catastrophic outliers by defining the bias b_z as the median rather than mean value of e_z. The catastrophic outlier rate $f_{\rm out}$ is defined as the fraction of galaxies with e_z greater than $\max(3\sigma_{\rm IQR}, 0.06)$.

For reference, Section 3.8 of the LSST Science Book (Abell et al. 2009) uses the standard definitions of these parameters in requiring

- RMS scatter $\sigma < 0.02(1 + z_{\rm true})$
- bias $b_z < 0.003$
- catastrophic outlier rate $f_{\rm out} < 10$ per cent .

B3 Comparison of photo-z point estimate metrics

Figure B1 shows both point estimates for all codes 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 points trace the details of the catastrophic outlier populations.

The finite grid spacing of the photo-z PDFs induces some discretization in z_{PEAK}. The features perpendicular

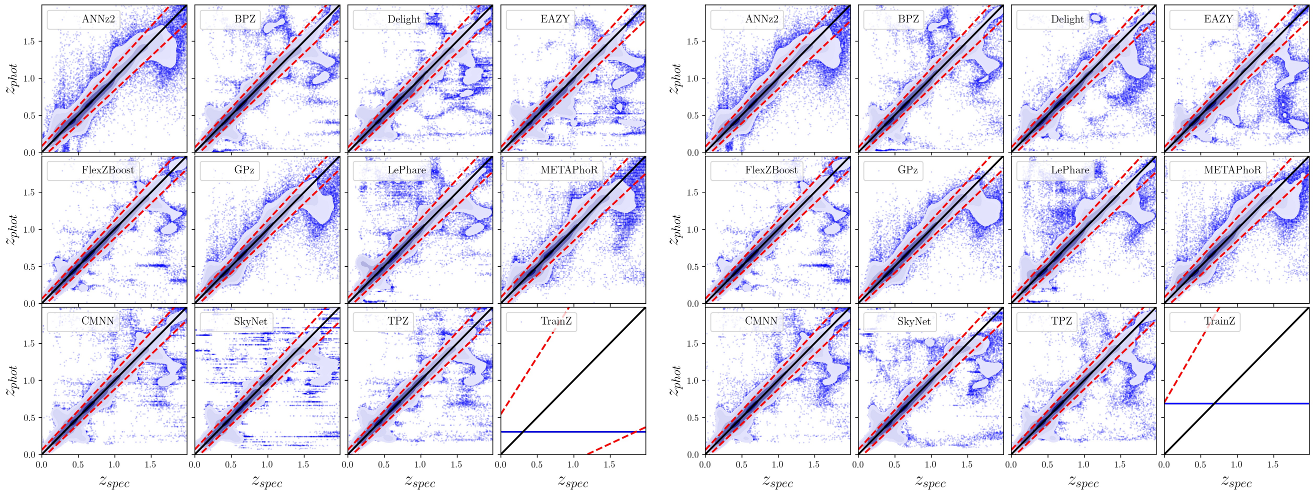

Figure B1. The density of photo-z point estimates (contours) reduced from the photo-z PDFs with outliers (blue) beyond the outlier cutoff (red dashed lines), via the mode (z_{PEAK}, left panel) and main-peak-mean (z_{WEIGHT}, right panel). The `trainZ` estimator (lower right sub-panels) has a shared z_{PEAK} and z_{WEIGHT} for the entire test set galaxy sample.

to the $z_{phot} = z_{spec}$ line are due to the 4000 Å break passing through the gaps between adjacent filters. Even the strongest codes feature populations far from the $z_{phot} = z_{spec}$ line representing a degeneracy in the space of colours and redshifts.

The intrinsic scatter, bias, and catastrophic outlier rate are given in Table B1. Perhaps unsurprisingly, performance under these metrics largely tracks that of the metrics of Section 4 of the photo-z PDFs from which the point estimates were derived. All twelve codes perform at or near the goals of the LSST Science Requirements Document[10] and Graham et al. (2018), which is encouraging if not unexpected for $i < 25.3$.

REFERENCES

Abbott T., et al., 2005, preprint (arXiv:astro-ph/0510346)
Abell P. A., et al., 2009, preprint (arXiv:0912.0201),
Aihara H., et al., 2018a, PASJ, 70, S4
Aihara H., et al., 2018b, PASJ, 70, S8
Almosallam I. A., Lindsay S. N., Jarvis M. J., Roberts S. J., 2016a, MNRAS, 455, 2387
Almosallam I. A., Jarvis M. J., Roberts S. J., 2016b, MNRAS, 462, 726
Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109
Benítez N., 2000, ApJ, 536, 571
Bernstein G., Huterer D., 2010, MNRAS, 401, 1399
Blanton M. R., Roweis S., 2007, AJ, 133, 734
Blanton M. R., et al., 2005, AJ, 129, 2562
Bonnett C., 2015, MNRAS, 449, 1043
Bonnett C., 2016, Python wrapper to SkyNet, `https://pyskynet.readthedocs.io/en/latest/`
Bonnett C., et al., 2016, Phys. Rev. D, 94, 042005
Bordoloi R., Lilly S. J., Amara A., 2010, MNRAS, 406, 881
Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503
Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees, Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A
Brescia M., Cavuoti S., Amaro V., Riccio G., Angora G., Vellucci C., Longo G., 2018, preprint, (`arXiv:1802.07683`)
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Carrasco Kind M., Brunner R. J., 2014, MNRAS, 442, 3380
Cavuoti S., Amaro V., Brescia M., Vellucci C., Tortora C., Longo G., 2017, MNRAS, 465, 1959
Chen T., Guestrin C., 2016, in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. ACM, New York, NY, USA, pp 785–794, doi:10.1145/2939672.2939785, `http://doi.acm.org/10.1145/2939672.2939785`
Connolly A. J., et al., 2014, in Angeli G. Z., Dierickx P., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 9150, Modeling, Systems Engineering, and Project Management for Astronomy VI. p. 14, doi:10.1117/12.2054953
Dahlen T., et al., 2013, ApJ, 775, 93
Dawson W. A., Schneider M. D., Tyson J. A., Jee M. J., 2016, ApJ, 816, 11
DeRose J., et al., 2019, arXiv e-prints, p. arXiv:1901.02401
Erben T., et al., 2013, MNRAS, 433, 2545
Fernández-Soto A., Lanzetta K. M., Yahil A., 1999, ApJ, 513, 34
Firth A. E., Lahav O., Somerville R. S., 2003, MNRAS, 339, 1195
Freeman P. E., Izbicki R., Lee A. B., 2017, MNRAS, 468, 4556
Graff P., Feroz F., Hobson M. P., Lasenby A., 2014, MNRAS, 441, 1741
Graham M. L., Connolly A. J., Ivezić Ž., Schmidt S. J., Jones R. L., Jurić M., Daniel S. F., Yoachim P., 2018, AJ, 155, 1
Green J., et al., 2012, preprint (arXiv:1208.4012),
Hildebrandt H., et al., 2010, A&A, 523, A31
Hofmann B., Mathé P., 2018, Inverse Problems, 34, 015007
Hunter J. D., 2007, Matplotlib: A 2D Graphics Environment, doi:10.1109/MCSE.2007.55
Ilbert O., et al., 2006, A&A, 457, 841
Ivezić Ž., et al., 2008, preprint (arXiv:0805.2366),
Izbicki R., Lee A. B., 2017, Electron. J. Statist., 11, 2800
Izbicki R., Lee A. B., Freeman P. E., 2017, Ann. Appl. Stat., 11, 698
Laureijs R., et al., 2011, preprint (1110.3193),
Leistedt B., Hogg D. W., 2017, ApJ, 838, 5

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

Table B1. Photo-z point estimate statistics

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

Malz A., Hogg D., in prep., CHIPPR, chippr, `https://github.com/aimalz/chippr`
Malz A., Marshall P., 2018, qp: Quantile parametrization for probability distribution functions (ascl:1809.011)
Malz A. I., Marshall P. J., DeRose J., Graham M. L., Schmidt S. J., Wechsler R., (LSST Dark Energy Science Collaboration 2018, AJ, 156, 35
Mandelbaum R., et al., 2008, MNRAS, 386, 781
Massarotti M., Iovino A., Buzzoni A., 2001, A&A, 368, 74
Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes J. D., Castander F. J., Paltani S., 2017, ApJ, 841, 111
Newman J. A., 2008, ApJ, 684, 88
Newman J. A., et al., 2015, Astroparticle Physics, 63, 81
Oliphant T., 2007, Python for Scientific Computing, doi:10.1109/MCSE.2007.58
Oyaizu H., Lima M., Cunha C. E., Lin H., Frieman J., 2008, ApJ, 689, 709
Polsterer K. L., D'Isanto A., Gieseke F., 2016, preprint (arXiv:1608.08016),
Rasmussen C., Williams C., 2006, Gaussian Processes for Machine Learning. Adaptative computation and machine learning series, MIT Press, Cambridge, MA
Rau M. M., Seitz S., Brimioulle F., Frank E., Friedrich O., Gruen D., Hoyle B., 2015, MNRAS, 452, 3710
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, ApJ, 771, 30
Sadeh I., Abdalla F. B., Lahav O., 2016, PASP, 128, 104502
Sánchez C., et al., 2014, MNRAS, 445, 1482
Schmidt M., 2005, minFunc: Unconstrained Differentiable Multivariate Optimization in Matlab, `http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`
Scott D. W., 1992, Multivariate Density Estimation. Theory, Practice, and Visualization. Wiley
Sheldon E. S., Cunha C. E., Mandelbaum R., Brinkmann J., Weaver B. A., 2012, The Astrophysical Journal Supplement Series, 201, 32
Skrutskie M. F., et al., 2006, AJ, 131, 1163
Tanaka M., et al., 2018, PASJ, 70, S9
The LSST Dark Energy Science Collaboration et al., 2018, preprint, (`arXiv:1809.01669`)
Waskom M., et al., 2017, doi:10.5281/zenodo.824567
York D. G., et al., 2000, AJ, 120, 1579
de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., 2013, Exp. Astron., 35, 25
de Jong J. T. A., et al., 2017, A&A, 604, A134