Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **524,** 5109–5131 (2023) https://doi.org/10.1093/mnras/stad1962
Advance Access publication 2023 July 10

Weak lensing tomographic redshift distribution inference for the Hyper Suprime-Cam Subaru Strategic Program three-year shape catalogue

Markus Michael Rau [1,2★] Roohi Dalal [3] Tianqing Zhang [1] Xiangchong Li [1] Atsushi J. Nishizawa [4,5,6] Surhud More,[7,8] Rachel Mandelbaum [1] Hironao Miyatake [8,5,6] Michael A. Strauss[3] and Masahiro Takada[8]

[1]*McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[2]*High Energy Physics Division, Argonne National Laboratory, Lemont, IL 60439, USA*
[3]*Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA*
[4]*Digital Transformation (DX) Center, Gifu Shotoku Gakuen University, Gifu 501-6194, Japan*
[5]*Institute for Advanced Research, Nagoya University, Nagoya 464-8602, Japan*
[6]*Kobayashi-Maskawa Institute for the Origin of Particles and the Universe (KMI), Nagoya University, Nagoya 464-8602, Japan*
[7]*Inter University Centre for Astronomy and Astrophysics, PB 4, Ganeshkhind, Pune 411007, India*
[8]*Kavli Institute for the Physics and Mathematics of the Universe (WPI), The University of Tokyo Institutes for Advanced Study (UTIAS), The University of Tokyo, Chiba 277-8583, Japan*

Accepted 2023 June 10. Received 2023 June 10; in original form 2022 November 29

ABSTRACT

We present posterior sample redshift distributions for the Hyper Suprime-Cam Subaru Strategic Program Weak Lensing three-year (HSC Y3) analysis. Using the galaxies' photometry and spatial cross-correlations, we conduct a combined Bayesian Hierarchical Inference of the sample redshift distributions. The spatial cross-correlations are derived using a subsample of Luminous Red Galaxies (LRGs) with accurate redshift information available up to a photometric redshift of $z < 1.2$. We derive the photometry-based constraints using a combination of two empirical techniques calibrated on spectroscopic and multiband photometric data that cover a spatial subset of the shear catalogue. The limited spatial coverage induces a cosmic variance error budget that we include in the inference. Our cross-correlation analysis models the photometric redshift error of the LRGs to correct for systematic biases and statistical uncertainties. We demonstrate consistency between the sample redshift distributions derived using the spatial cross-correlations, the photometry, and the posterior of the combined analysis. Based on this assessment, we recommend conservative priors for sample redshift distributions of tomographic bins used in the three-year cosmological Weak Lensing analyses.

Key words: methods: data analysis – methods: numerical – methods: statistical – techniques: photometric – galaxies: distances and redshifts – cosmology: observations.

1 INTRODUCTION

Cosmological weak lensing (WL) and structure growth analyses for the current and next generation of large area photometric surveys like the Dark Energy Survey (DES; e.g. Abbott et al. 2018), the Kilo-Degree Survey (KiDS; e.g. Hildebrandt et al. 2017), the Hyper Suprime-Cam (HSC; e.g. Aihara et al. 2018), the Rubin Observatory Legacy Survey of Space and Time (LSST; e.g. Ivezić et al. 2019), the Roman Space Telescope (e.g. Spergel et al. 2015), and Euclid (e.g. Laureijs et al. 2011) depend on accurately accounting for sources of systematic bias and uncertainty (e.g. Mandelbaum 2018). The primary cosmological probes in these campaigns are measurements of the growth of structure based on two-point statistics of galaxy and gravitational shear fields (see e.g. Hikage et al. 2019; Hamana et al. 2020; Asgari et al. 2021; Giblin et al. 2021; Heymans et al. 2021; Joachimi et al. 2021; Abbott et al. 2022; Amon et al. 2022; Pandey et al. 2022; Prat et al. 2022; Secco et al. 2022a).

Since measurements of the broad-band photometry of galaxies only allow us to extract limited redshift information, measurements of two-point statistics of density fields are typically considered in projection along the line of sight. The line of sight or sample redshift distribution $p_{\rm samp}(z)$ enters the corresponding WL and Large-Scale Structure (LSS) theory predictions, which are used to constrain cosmological parameters using measurements of the projected density fields in a likelihood framework. In order to calibrate the credible intervals on cosmological parameters, it is important to characterize and control sources of systematic bias and uncertainty in $p_{\rm samp}(z)$ estimates (see e.g. Huterer et al. 2006; Hoyle et al. 2018; Tanaka et al. 2018; Hikage et al. 2019; Joudaki et al. 2020; Hildebrandt et al. 2021).

★ E-mail: markusmichael.rau@googlemail.com

One primary science driver for photometric surveys is to constrain the dark energy equation-of-state parameters by measuring the distance–redshift and growth–redshift relations (see e.g. Albrecht et al. 2006, p. 31) which both enter the WL and LSS modelling and parametrize the growth of structure and expansion history of our universe. This approach leads to degeneracies between cosmological parameters that describe the cosmic density fields, $p_{\rm samp}(z)$ parameters that enter the aforementioned line-of-sight projection kernel (e.g. Ma, Hu & Huterer 2006; Bernstein & Huterer 2010), and other modelling components such as the galaxy-dark matter bias (see e.g. Matarrese et al. 1997; Clerkin et al. 2015; Chang et al. 2016; Prat et al. 2018; Simon & Hilbert 2018; Sugiyama et al. 2020; Stölzner et al. 2022) and intrinsic alignments (Amon et al. 2022; Sánchez et al. 2022; Secco et al. 2022b). Parameters that describe the sample redshift distribution for samples of galaxies can therefore exhibit a degeneracy with cosmological or astrophysical parameters. Inaccuracies in the distance (or redshift) measurements of ensembles of galaxies are therefore important for modelling systematics in these surveys.

The two main sources of information available to constrain redshifts of individual galaxies as well as samples of galaxies are measurements of their photometry and spatial clustering. Methods that exploit photometric information (for a recent review, see Salvato, Ilbert & Hoyle 2019; Newman & Gruen 2022) can be broadly categorized into two classes. Empirical methods (Tagliaferri et al. 2003; Collister & Lahav 2004; Gerdes et al. 2010; Carrasco Kind & Brunner 2013; Bonnett 2015; Rau et al. 2015; Hoyle 2016) utilize calibration data to directly learn a mapping from the measured photometry to the redshift of galaxies given a spectroscopic survey. Template fitting methods (e.g. Arnouts et al. 1999; Benítez 2000; Feldmann et al. 2006; Ilbert et al. 2006; Greisel et al. 2015; Leistedt, Mortlock & Peiris 2016; Malz & Hogg 2020) use a forward model that constrains the redshift of galaxies using a likelihood of the 'reproduced' galaxy flux, given a model for the galaxy spectral energy distribution (SED) and other parameters of interest.

Both of these approaches lead to consistent estimators if their underlying assumptions are met and a correct statistical estimator is constructed. However, in real data, incorrectly modelled selection functions and modelling uncertainties can lead to significant model misspecification. A particular example are selection functions in spectroscopic data sets used for redshift calibration (Masters et al. 2017, 2019; Hartley et al. 2020), due to the impractically long exposure times required to spectroscopically observe colour-complete samples at faint magnitudes (see e.g. Huterer et al. 2014; Newman et al. 2015). One goal of this paper is to discuss and discern the assumptions made in various $p_{\rm samp}(z)$ inference methodologies by discussing them in a unified likelihood framework.

As mentioned, a second method to constrain $p_{\rm samp}(z)$ are spatial cross-correlations between photometric and spectroscopic samples (e.g. Newman 2008; McQuinn & White 2013; Ménard et al. 2013; Scottez et al. 2016; Davis et al. 2017; Morrison et al. 2017; Raccanelli, Rahman & Kovetz 2017; Gatti et al. 2018; van den Busch et al. 2020; Hildebrandt et al. 2021). Since the photometric and spectroscopic samples trace the same underlying dark-matter field, the amplitudes of the two-point function measured between spectroscopic samples (binned in redshift) and the full photometric sample (with no accurate redshift information) can constrain the sample redshift distribution of the full photometric sample $p_{\rm samp}(z)$. Redshift-dependent galaxy-dark matter bias of the photometric and spectroscopic samples, cosmic magnification effects (see e.g. Scranton et al. 2005), and the redshift evolution of the underlying dark-matter density field affect the aforementioned relative redshift bin heights.

While it is a challenge to correct for these degenerate effects, cross-correlations are one of the most important techniques for $p_{\rm samp}(z)$ calibration today. We note that two-point statistics from e.g. WL (e.g. Benjamin et al. 2013; Stölzner et al. 2021), or shear-ratios (e.g. Prat et al. 2019; Giblin et al. 2021; Sánchez et al. , 2022) can also be used in the context of redshift estimation.

However, since WL in particular is considered one of the most promising methods to constrain dark energy, photometric redshift estimation is treated in our analysis as a systematic that enters the theoretical modelling of a separate 'cosmological' likelihood rather than using WL statistics as a redshift estimation technique. Recently, the question of how to integrate redshift uncertainty into a likelihood of two-point statistics has been considered (McLeod, Balan & Abdalla 2017; Hoyle & Rau 2019), especially in the context of how to combine template fitting and cross-correlation measurements (Jones & Heavens 2019; Sánchez & Bernstein 2019; Alarcon et al. 2020; Rau, Wilson & Mandelbaum 2020; Myles et al. 2021; Cawthon et al. 2022; Gatti et al. 2022; Rau et al. 2022; Zhang et al. 2023). In Rau et al. (2022), we developed a Bayesian hierarchical inference framework that self-consistently combines information from both cross-correlation redshift estimation and photometry, specifically discussing aspects of regularization and probability calibration. Rau et al. (2022) validate the basic aspects of our presented methodology using mock data where well-controlled sources of systematics are modelled. While the usage of simulated mock data necessarily has limitations, we performed this analysis with the greatest possible realism in mind. We found that a hierarchical modelling approach similar to the one presented in this paper can indeed reach the level of accuracy necessary for LSST, as measured using common performance metrics.

This paper presents the sample redshift inference methodology for the HSC Y3 cosmological WL analysis, which consists of two cosmic shear analyses (Dalal et al. 2023; Li et al. 2023a) in four tomographic bins and a 3x2pt analysis (Miyatake et al. 2023; More et al. 2023; Sugiyama et al. 2023) which uses one tomographic bin. This paper presents our inference methodology in the context of the cosmic shear analyses, where it was used as the default method for redshift inference. Tomography refers here to binning the shear catalogue along the redshift dimension, using a predictor for redshift. While the separation of these tomographic samples in redshift is typically not perfect, i.e. the sample redshift distributions of adjacent tomographic bins will overlap, autocorrelations and cross-correlations estimated on the tomographic samples will have more information about the redshift evolution of the growth of structure than the two-point function estimated on the unbinned sample. We utilize five band photometry in the *grizy* filter set to infer the tomographic sample redshift distributions (tomographic $p_{\rm samp}(z)$). We apply our methodology to the Hyper Suprime-Cam three-year shape catalogue[1] data set (HSC Y3), and derive and recommend prior distributions over a $p_{\rm samp}(z)$ parametrization that can be used in the subsequent cosmological WL analyses.

This work presents a significant update to the HSC sample redshift inference methodology developed for the first year (HSC Y1) analyses presented in Hikage et al. (2019) and Hamana et al. (2020). This is vital, since the increased area of the shear catalogue from 136.9 deg^2 (HSC Y1) to 433.5 deg^2 (HSC Y3) implies that our redshift calibration accuracy has to significantly improve to prevent

[1]Data observed through 2019.

systematic biases or uncertainties in cosmological parameters from dominating over the statistical uncertainties.

2 MOTIVATION

The HSC Y1 sample redshift distribution calibration described in Hikage et al. (2019) and applied in the context of the Y3 cosmic shear analysis in that work and in Hamana et al. (2020) estimates the sample redshift distributions in tomographic bins by reweighting COSMOS2015 (Ilbert et al. 2006; Laigle et al. 2016) galaxies in colour space. The quantification of uncertainty includes a systematic error budget derived by comparing the reweighted sample redshift distribution with the sample redshift distribution estimators obtained from a set of seven independent methods. The HSC Y1 analyses used uncertainties in the means of the tomographic redshift distributions as parameters to marginalize over photometric redshift uncertainty.

The forthcoming HSC Y3 analyses also include a systematic error budget based on a comparison of models, but presents a significantly updated framework for sample redshift inference that includes a treatment of cosmic variance as well as a cross-correlation calibration of sample redshift distributions based on a sample of Luminous Red Galaxies (LRGs, Oguri 2014; Oguri et al. 2018a, b; Ishikawa et al. 2021) selected using the Cluster finding algorithm based on Multi-band Identification of Red-sequence gAlaxies (CAMIRA). We will abbreviate this sample as 'CAMIRA LRG' in the following. The inclusion of a cross-correlation data vector into the inference of the sample redshift distribution $p_{\rm samp}(z)$ is arguably the most significant improvement over the Y1 analyses, as it allows us to independently test the quality of the estimated $p_{\rm samp}(z)$ in the tomographic bins.

We refer to the remainder of the paper for an explanation of the HSC Y3 redshift inference methodology. However, we would like to motivate the effect that these significant changes have on our redshift calibration using a forecast, which is based on a mock Y3 WL cosmological analysis. We perform a mock analysis of a synthetic data vector with a redshift distribution inferred for the HSC Y1 shape catalogue, using a similar analysis to the one presented in this work, and compare it with an analysis using the simple 'stacked' redshift distribution from Hamana et al. (2020). The main difference from the methodology described in the rest of this work is the usage of the Dirichlet distribution, as well as the usage of a model combination scheme described in Rau et al (in preparation) that accounts for the model uncertainty across the several different photometric redshift codes applied to the HSC Y1 data set and described in Hamana et al. (2020).

The sample redshift posteriors and the inference scheme employed to marginalize over the uncertainty in those parameters are both described in Zhang et al. (2023). The sampling is based on using the mean of the tomographic redshift distributions as the main parameter over which we marginalize (referred to as the 'shift model' in the following). The cosmological parameter inference is performed using the multinest method with 500 live points. We consider nine cosmological and nine astrophysical parameters and four parameters within the shift model. The simulated data vector includes noise based on the scaled HSC first year covariance as described in Zhang et al. (2023). Both contours shown in Fig. 1 use the shift model to marginalize over the $p_{\rm samp}(z)$ uncertainties, where our prior on the tomographic $p_{\rm samp}(z)$ is generated using the mean redshifts of 1000 samples of sample redshift posterior generated by the updated methodology. The prior on the mean redshift of the stacked redshift distribution follows Hamana et al. (2020). We generate an approximation to the Y3 covariance by dividing the Y1 covariance by 3, which approximately accounts for the increase in area from Y1 to Y3 while ignoring changes in the contiguity of the survey footprint. Fig. 1 compares the posteriors in the $\Omega_m - S_8$ plane.

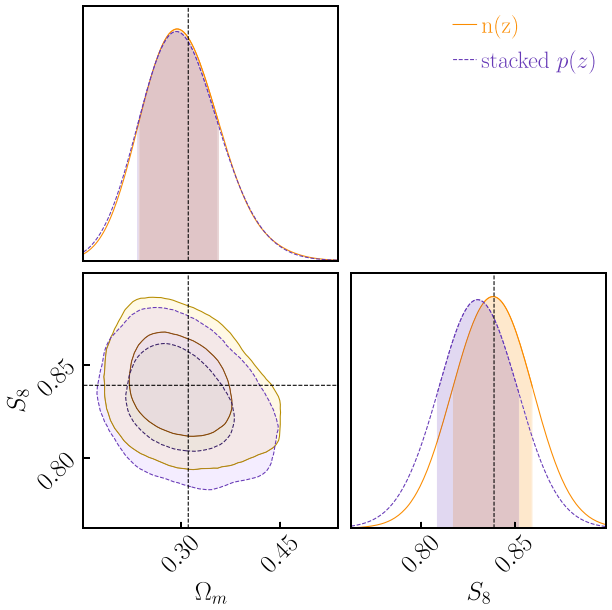

Figure 1. Forecast of the impact and importance of using an updated $p_{\rm samp}(z)$ inference methodology on cosmological inference from the predicted Y3 data vector and covariance matrix. The purple contour uses the 'stacked' redshift distribution for the Y1 galaxy catalogue, while the orange contour uses the Y1 redshift distribution inferred from an analysis similar to this work. The change in redshift distribution causes a 0.5σ shift in the S_8 constraints, which is significant for the upcoming Y3 cosmic shear analyses.

We note a 0.5σ shift in S_8, which shows that the updated analysis would predict a higher S_8 value. Note that the synthetic data vector is generated with the updated redshift distribution, so the analysis with that redshift distribution recovers the true cosmological parameters. This figure illustrates the importance of $p_{\rm samp}(z)$ calibration and in particular of a joint $p_{\rm samp}(z)$ analysis that includes complementary data sources and analysis techniques.

3 DATA

The following sections describe the data sets and catalogues that we use in this work. Specifically, we consider three data sets that are relevant at different stages of the analysis. Section 3.1 describes the photometric data included in the HSC shear catalogue, Section 3.2 the catalogue of LRGs (Oguri 2014; Oguri et al. 2018a, b; Ishikawa et al. 2021) that we will use for our cross-correlation analysis, and Section 3.3 a matched catalogue between the photometric data and spatially overlapping spectroscopic surveys. We will abbreviate the photometric data included in the HSC shear catalogue as 'HSC phot', the catalogue of LRGs as 'CAMIRA LRG', and the matched catalogue as 'specXphot'.

3.1 HSC Y3 shape catalogue

The Hyper Suprime-Cam survey, which is part of the Subaru Strategic Program (SSP), is an optical imaging survey carried out using the Hy-

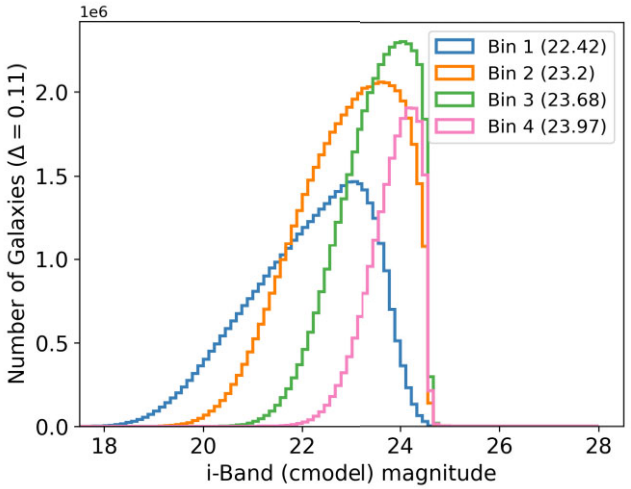

Figure 2. Distribution of *i*-band cmodel magnitudes for the four tomographic bins. We show the *i*-band cmodel magnitudes on the horizontal axis and the number of galaxies on the *y*-axis. The median magnitudes are shown in the legend, the magnitude bin size is $\Delta = 0.11$.

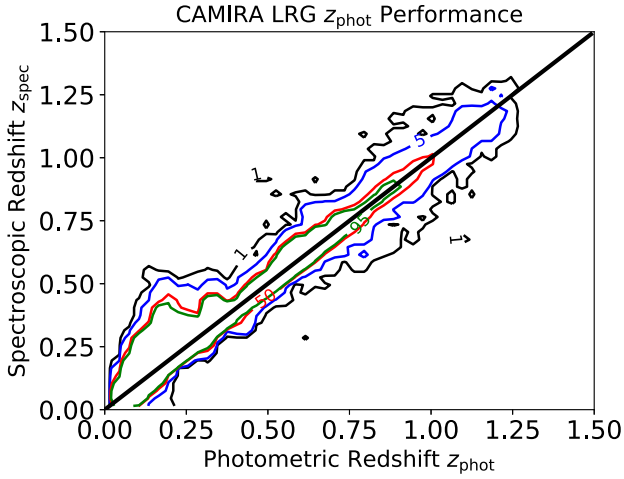

Figure 3. Photometric versus spectroscopic redshift for CAMIRA LRG galaxies with spectroscopic redshifts. The dashed black line denotes perfect photometric redshift prediction. There is a small population (0.02 per cent) of redshift outliers at $z_{\rm spec} > 5$ which we do not show here. The contour line annotations indicate the corresponding probability density values in per cent.

per Suprime-Cam (HSC, Miyazaki et al. 2018), a wide field camera with 1.77 deg^2 field of view installed on the 8.2 m Subaru telescope. The shear catalogue we use in this work, as part of the year-3 analysis, consists of 417 deg^2[2] of wide-field optical galaxy photometry in *grizy* with a 5σ limiting magnitude of $r \approx 26$. We refer the reader to Aihara et al. (2018) and Aihara et al. (2022) for a more detailed overview of the design of the HSC survey. The catalogues from this internal data release along with the shape catalogue and their calibrations are expected to be made public as part of a future incremental update to PDR3 (Aihara et al. 2022) after the cosmological analyses are finished.

Fig. 2 plots the cmodel[3] magnitude distribution in the *i* band for the four tomographic bins. The tomographic bins ('Bin 1', 'Bin 2', 'Bin 3', 'Bin 4') are selected using a procedure described in Section 5.2 to have approximately the redshift ranges of (0.3, 0.6], (0.6, 0.9], (0.9, 1.2], and (1.2, 1.5].

We see that all four tomographic bins extend to magnitudes fainter than 24 in the *i* band, where the majority of galaxies have a magnitude around that value. Bins 1–4 contain 24, 33, 28, and 15 per cent of the galaxies, respectively, and the raw (effective) galaxy number densities are 3.92 (3.77), 5.63 (5.07), 4.68 (4.00), and 2.60 (2.12) arcmin^{-2}. Since we present this analysis in the context of the upcoming cosmic shear analysis for HSC Y3, we apply our methodology to galaxies contained in the shear catalogue that has a magnitude limit of 24.5. We therefore need to include all of the lensing selection criteria and lensing weights throughout the analysis. Lensing weights are inverse variance weights derived in the construction of the galaxy shape estimate. For a description of the methodology to derive these selection criteria and lensing weights, we refer to Li et al. (2022). In the following text, we will refer to the shear catalogue as 'HSC phot'.

3.2 CAMIRA LRG sample

The CAMIRA LRG sample[4] contains LRGs selected using the CAMIRA algorithm (Oguri 2014; Oguri et al. 2018b; Ishikawa et al. 2021). CAMIRA identifies LRGs as red-sequence galaxies based on their photometry and their consistency with the expected colours from stellar population synthesis models. The LRG sample has a limited redshift range of $z < 1.2$ and the redshifts of these LRGs are subject to photometric redshift error.[5] In this work, we use the CAMIRA LRG sample as a reference catalogue for spatial cross-correlations with galaxy samples from HSC phot. This will allow us to construct a likelihood that constrains the $p_{\rm samp}(z)$. Since the LRG galaxy population provides a photometric sample with good redshift quality and well-understood clustering properties, it is the ideal reference sample for cross-correlation studies. However, as we will describe in Section 5.5, we need to marginalize over the photometric redshift error of the LRGs. This requires a model for the photometric redshift error of the CAMIRA LRG galaxies, which we detail there. The photometric redshift error model is calibrated using the corresponding LRG subsample of the full specXphot reference sample described in Section 3.3. Fig. 3 shows the photometric redshift of the CAMIRA LRGs against the spectroscopic redshifts of the aforementioned specXphot reference subsample as a contour plot. We see that, especially around $z_{\rm spec} \approx 0.4/z_{\rm phot} \approx 0.2$, a well-known redshift region where the 4000 Å break crosses between the *g* and the *r* filters, the photometric redshift of the CAMIRA LRG galaxies shows a mean bias in the contour lines, although we identify a small number of outlier galaxies with $z_{\rm spec} > 5$. This population consists of 0.02 per cent of the full CAMIRA LRG specXphot reference sample;

[2]We remove a 20 deg^2 region that failed the cosmic shear B-mode test (see Zhang et al. 2023).
[3]The SDSS CModel magnitude (Lupton et al. 2001; Abazajian et al. 2004) algorithm fits a galaxy using elliptical models with both an exponential profile and a de Vaucouleurs profile. The derived CModel flux is approximately a linear interpolation between exponential and de Vaucouleurs models. We refer to Huang et al. (2017) for more details.

[4]https://github.com/oguri/cluster_catalogs/tree/main/hsc_s20a_camira (Accessed 2022 June 10)
[5]Photometric redshifts for LRGs are often derived using SED fitting techniques and have significantly better redshift accuracy compared with the full photometric sample.

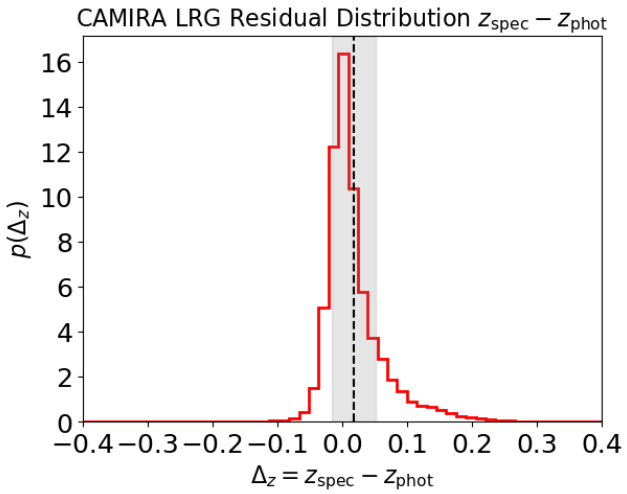

Figure 4. Distribution of photometric redshift residuals of $z_{\rm spec} - z_{\rm PhotZ}$. The black dashed vertical line denotes the mean, while the grey contours show the range between the 16th and 84th percentiles (selected to resemble a 'Gaussianized' 1σ interval).

the contamination is small and we leave a further investigation of the outlier population for future work. The bias at low photometric redshift is also apparent in the right tail of Fig. 4, which shows a histogram of the residual redshift error $z_{\rm spec} - z_{\rm Phot}$. The black dashed vertical line shows the mean residual redshift error (0.018), while the grey region visualizes the range between the 16th (-0.017) and 84th (0.052) percentiles (equivalent to the 'Gaussian' $\pm 1\sigma$ intervals).

3.3 Spectroscopic reference samples

This section gives an overview of the spectroscopic reference samples that are available to match against HSC phot to generate the 'specXphot' calibration sample. We will concentrate on the aspects that are relevant for this work and refer to Tanaka et al. (2018) for a more detailed description of the reference samples and the selection criteria used to generate them. The reference sample (Nishizawa et al. 2020) is assembled from the following sources: zCOSMOS DR3 (Lilly et al. 2009), zCOSMOS faint (Lilly et al. 2009) including private spectroscopic data[6], COSMOS2015 (Laigle et al. 2016), UDSz (Bradshaw et al. 2013; McLure et al. 2013), 3D-HST (Skelton et al. 2014; Momcheva et al. 2016), FMOS-COSMOS (Silverman et al. 2015), VVDS (Le Fèvre et al. 2013), VIPERS PDR1 (Garilli et al. 2014), SDSS DR12 (Alam et al. 2015), GAMA DR2 (Liske et al. 2015), WiggleZ DR1 (Drinkwater et al. 2010), DEEP2 DR4 (Davis et al. 2003; Newman et al. 2013), VANDELS DR2 (Pentericci et al. 2018), C3R2 (Masters et al. 2017, 2019), and PRIMUS DR1 (Coil et al. 2011; Cool et al. 2013). The spectroscopic redshift measurements are extracted from both high-quality spectroscopic measurements (≈170 000 galaxies) and lower resolution prism spectroscopy (≈37 000 galaxies). In addition, Tanaka et al. (2018) also include 170 000 Cosmos2015 multiband photometric redshifts and a sample of privately obtained spectroscopic redshifts (Mara Salvato private communication).

Tanaka et al. (2018) homogenize the catalogue to ensure approximately uniform data quality. This is done by imposing cuts on the quality flags in the respective source catalogues. The galaxies are then matched to HSC phot (see Section 3.1) to create the specXphot reference sample. This catalogue contains both the photometric measurements in HSC phot and the spectroscopic redshift estimates from the listed sources.

We will utilize this data set as a reference sample to calibrate and train photometric redshift estimates. While the selection cuts imposed by Tanaka et al. (2018) are designed to minimize the impact of colour-redshift incompleteness on photometric redshift estimates trained on the specXphot calibration sample, we still have to consider the spatial selection function due to the much smaller survey footprint of the specXphot sample in relation to HSC phot. Furthermore, residual selection function induced systematics will likely remain, which motivates our usage of cross-correlations for redshift calibration.

To give an overview of this data set, Fig. 5 shows the normalized spectroscopic redshift distribution of the specXphot sample (upper panel), the histogram of the i-band magnitude (middle panel), and the spatial area covered by the specXphot calibration catalogue up to (i.e. fainter than) the magnitude limit plotted on the horizontal axis (lower panel). The middle panel shows that the specXphot calibration catalogue covers the magnitude range of the HSC phot sample (black dashed histogram). We generate the lower panel by adding up the area as a function of i-band magnitude covered by the specXphot calibration catalogue using a healpix pixelization (Górski et al. 2005) with NSIDE = 1024. The black dashed horizontal line shows the size of the COSMOS2015 calibration field ($\approx 2\,{\rm deg}^2$) that dominates the data at the faint end. It represents the lower limit on the HSC Y3 area, for which we have available calibration data. This lower limit will be used in Section 5.4 to derive a conservative assessment of the cosmic variance error budget in our $p_{\rm samp}(z)$ inference methodology.

4 THE PHOTOMETRIC REDSHIFT PROBLEM

The $p_{\rm samp}(z)$ of galaxies is a vital component in the modelling of projected density fields in weak gravitational lensing and LSS. This one-point density distribution along the line of sight enters the projection kernel in the modelling of these probes. In this section, we summarize the foundational methodology for estimating the redshift distributions of galaxy ensembles ('$p_{\rm samp}(z)$ inference', hereafter).

There are two main approaches to the photometric redshift problem. The 'forward-modelling' approach models the data generating process[7] and treats the $p_{\rm samp}(z)$ as the prior on the redshift of individual galaxies. We note that 'traditional' approaches like SED fitting would also fall under this category. The alternative 'conditional density estimation' approach constructs a direct probabilistic mapping between the photometry of galaxies and their redshift. For HSC, we consider both methodologies, and therefore describe $p_{\rm samp}(z)$ inference in both scenarios in the following two subsections. We note however that the models that we select for our final inference ('DNNz' and 'DEMPz', see Section 5.1) are both conditional density estimation techniques. We still describe both methodologies in detail for completeness.

[6]Mara Salvato (private communication).

[7]The 'data generating process' refers to the procedure of drawing galaxy properties from population distributions like the sample redshift distribution and mapping these quantities to measured observables, like e.g. the photometry, via a likelihood (or sampling distribution).

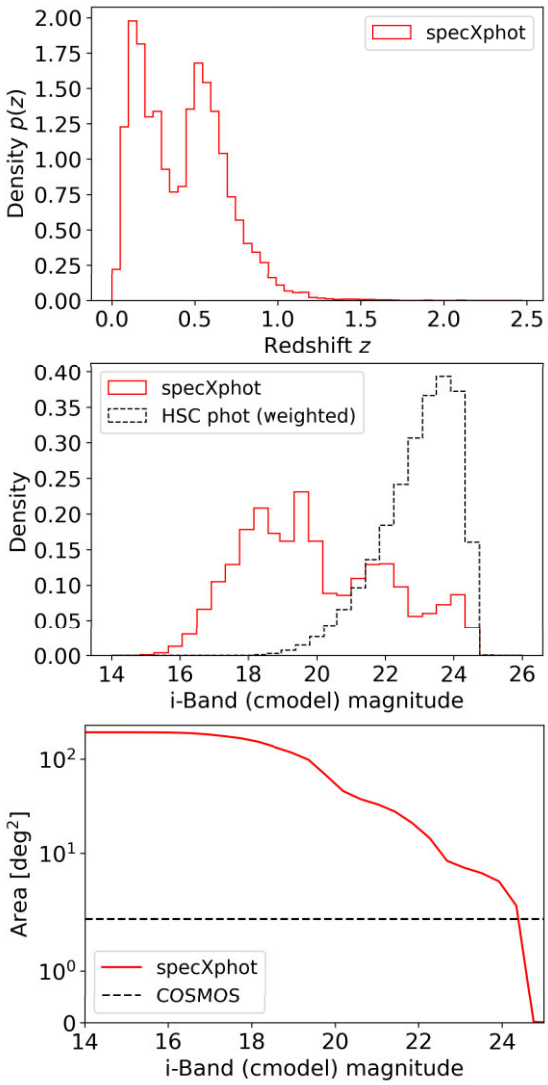

Figure 5. Illustration of the spatial coverage and the distribution of galaxies as a function of *i*-band magnitude for the specXphot Calibration data set used for $p_{\rm indiv}(z)$ estimation. Top: Spectroscopic redshift distribution of the specXphot calibration sample. The histogram is normalized to integrate to unity. Middle: Distribution of galaxies in *i*-band magnitude for the specXphot Calibration data set (red solid) and the HSC phot data set (black dashed) including lensing weights. Lower: Area in square degrees covered by the specXphot data set as a function of *i*-band magnitude. The vertical axis, plotted on the symmetrical log scale, shows the total area covered by all galaxies with *i*-band magnitude brighter than the value shown on the horizontal axis. The dashed horizontal line shows the area covered by the COSMOS2015 data set that dominates the specXphot data set at the faint end.

Throughout this paper we parametrize the $p_{\rm samp}(z)$ using a histogram with height parameters $\boldsymbol{\phi}_{\rm nz}$ for $N_{\rm bins}$ histogram bins as

$$p_{\rm samp}(z) = \sum_{i=1}^{N_{\rm bins}} \phi_{{\rm nz},i} \mathbb{1}_i(z)\,, \tag{1}$$

where $\mathbb{1}_i$ denotes the 'indicator' function for a given histogram bin *i*. The indicator function $\mathbb{1}(z)$ is unity if z falls in the histogram bin, and zero otherwise. We note that instead of a histogram parametrization one could also consider a kernel ansatz using, e.g. a Gaussian kernel. This could have advantages because we could consider a continuous approximation with (potentially) fewer parameters. However, this is not expected to be a vital reduction in approximation error. In the current analysis we decided to use the histogram, a flexible parametrization that does not necessitate the development of a specialized model for the $p_{\rm samp}(z)$. In the following subsections we will describe two methodologies to infer sample redshift distributions $p_{\rm samp}(z)$.

We want to briefly (and somewhat colloquially) comment on the different interpretation of $p_{\rm samp}(z)$ in both contexts. Both techniques formulate a likelihood for the parameters $\boldsymbol{\phi}$. The likelihood formulated in Section 4.1 describes a sampling distribution over the observed flux. The approach Section 4.2 describes a sampling distribution over parameters of a density estimate constructed using conditional density estimates that map directly from observed photometry to galaxy redshift. We highlight that it is important to distinguish both approaches and continue with a detailed description of each in the following subsections.

4.1 Forward-modelling approach

The goal of the forward modelling approach in general and SED modelling in particular is to formulate a statistical procedure that hierarchically models the relation between ensemble distributions of quantities of interest like galaxy redshift, type, or stellar mass, the corresponding properties of individual galaxies and observables like photometry.

In a simplified model (focussing on the redshift z as the quantity of interest) we can formulate this as (e.g. Leistedt et al. 2016; Malz & Hogg 2020; Rau et al. 2022)

$$p(\hat{\mathbf{F}}|\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\Omega}) = \prod_{i=1}^{N_{\rm gal}} \int {\rm d}z_i\, \omega_i\, p(\mathbf{f}_i|z_i, \boldsymbol{\Omega}) p(z_i|\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\Omega})\,. \tag{2}$$

Here, $\hat{\mathbf{F}}$ denotes the set of fluxes of all $N_{\rm gal}$ galaxies in the sample, $\mathbf{f}_i$ (z_i) denotes the flux in a filter set (redshift) of the individual galaxy with index *i*, and $\boldsymbol{\Omega}$ denotes a set of auxiliary parameters that describe other galaxy properties such as galaxy type or stellar mass. The factor ω_i denotes the lensing weight for galaxy *i*. We note that bold symbols denote vector quantities. Equation (2) assumes that the flux and redshift of each galaxy are drawn independently of any other. To simplify the notation we will implicitly assume conditioning on $\boldsymbol{\Omega}$, but omit it from the notation in the following discussion. Effects like blending (MacCrann et al. 2022; Li et al. 2023b) break the aforementioned assumption of independence of the galaxy flux measurements. This requires either the formulation of a joint flux likelihood of sets of galaxies or a reformulation of the likelihood on the pixel level to facilitate a joint inference with photometry and shear. We do not expect this approximation to dominate the error budget for this analysis and refer to future work. Also, note that Li et al. (2022) explored the connection between redshift and shear calibration in the context of simulations devised to explore blending effects for HSC survey data, and have already folded this effect into our understanding of redshift-dependent shear calibration.

We identify the term $p(\mathbf{f}_i|z_i, \boldsymbol{\Omega})$ in equation (2) as the likelihood of the observed individual galaxy flux given redshift, and the term $p(z_i|\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\Omega})$ as the prior distribution of the galaxy redshifts given the parameters that describe the sample redshift distribution (see equation (1) for the definition of these parameters). This specifies a forward model, where the individual galaxy redshifts z_i are first 'drawn' from the sample redshift distribution, denoted by the prior $p(z_i|\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\Omega})$. The likelihood then relates the drawn galaxy redshifts z_i to the observed galaxy fluxes $\mathbf{f}_i$ via the likelihood function

$p(\mathbf{f}_i|z_i, \mathbf{\Omega})$. We note that the sample redshift distribution $p_{\rm samp}(z)$ is here conditional on both the parameters $\boldsymbol{\phi}_{\rm nz}$ that are used to construct the distribution, as well as auxillary parameters $\mathbf{\Omega}$ that describe other quantities of interest.

In the following text we present a toy model that illustrates some aspects of the forward model formulation in a more concise manner. We also refer to Meister (2009), Rau et al. (2022), and Padmanabhan et al. (2005) for similar introductions. Simplifying the problem and notation we can relate equation (2) to the linear model

$$\boldsymbol{\phi}_{\rm nz\,noisy} = \boldsymbol{K} \cdot \boldsymbol{\phi}_{\rm nz\,true} \tag{3}$$

by identifying $p(\hat{\mathbf{F}}|\boldsymbol{\phi}_{\rm nz}, \mathbf{\Omega})$ with a 'smeared-out' and observed vector $\boldsymbol{\phi}_{\rm nz\,noisy}$, the set of likelihoods $\{p(\mathbf{f}_i|z_i, \mathbf{\Omega})\,|\,0 < i < N_{\rm gal}\}$ with the matrix $\boldsymbol{K}$ and the sample redshift distribution $p(z_i|\boldsymbol{\phi}_{\rm nz}, \mathbf{\Omega})$ with a noiseless, or 'true', vector $\boldsymbol{\phi}_{\rm nz\,true}$.

Thus, to recover $\boldsymbol{\phi}_{\rm nz\,true}$ we need to invert the matrix $\boldsymbol{K}$, which can be very sensitive to small variations in $\boldsymbol{\phi}_{\rm nz\,noisy}$ or the matrix $\boldsymbol{K}$. The former could be caused, for example, by the photometric noise, the latter by model error in the forward model. The sensitivity of the linear model on these variations depends on the condition number of $\boldsymbol{K}$, which will in turn depend on the resolution of the reconstruction, i.e. the histogram width in our parametrization. The forward modelling approach therefore treats $p_{\rm samp}(z)$ inference as an inverse problem whose solution is critically dependent on accurate modelling of the individual galaxy likelihoods and the regularization strategies that we impose. The likelihood modelling should also include how galaxies are selected into tomographic bins and other selection functions.

Typically one needs to 'regularize' this inverse problem. Regularization techniques reduce the noise in the reconstructed $p_{\rm samp}(z)$ by adding constraints to its shape. Ideally this information is not chosen arbitrarily, but rather results from data-driven constraints (e.g. a cross-correlation data vector that is included into the inference). We refer to a more detailed discussion on regularization and its methodological challenges in our previous work (Rau et al. 2022). We would like to note that instead of analytically modelling the likelihood function, one can also impose a synthetic likelihood. This can be done for example using a density estimate constructed using a Self-Organizing Map (see e.g. Kohonen 1982) that is trained on calibration data as in e.g. Sánchez & Bernstein (2019), Alarcon et al. (2020), and Myles et al. (2021). In this case the same considerations would apply, where we can substitute the analytical likelihood with a likelihood that is empirically estimated. One of the methods considered but ultimately not selected in this work is the Mizuki SED fitting method (Tanaka 2015; Tanaka et al. 2018). Mizuki is an SED fitting technique that formulates an analytic likelihood function, so the techniques described in this section directly apply. In Appendix B, we provide a detailed description of our sample redshift inference methodology.

4.2 Conditional density estimation approach

The conditional density estimation approach (see e.g. Lima et al. 2008; Carrasco Kind & Brunner 2013; Rau et al. 2015; Dalmasso et al. 2020) constructs a density estimate between the photometry of galaxies and the redshift $p(z|\mathbf{f})$ using a calibration, or training, data set. As such, the conditional density estimation approach depends on the calibration data set to constrain the conditional distribution $p(z|\mathbf{f})$. The calibration data set provides information about the mapping between photometry and redshift and the probability density of redshift given photometry.

In contrast, forward modelling explicitly considers a likelihood function or, alternatively, constructs a sampling distribution using numerical simulations. The forward modelling approach therefore must include information on the relative abundance of galaxies of different type and redshift into the prior (or as part of the simulation draws). Imposing a prior on the population distributions such as the $p_{\rm samp}(z)$ effectively acts as a regularization.[8]

For the conditional density estimation approach, one can formulate an estimate for the sample redshift distribution via marginalization

$$p_{\rm samp}(z) = \int d\mathbf{f}\; p(z|\mathbf{f}) p(\mathbf{f})\,. \tag{4}$$

Equation (4) also describes a linear system, similar to equation (3). However, equation (4) is typically much better 'conditioned' than equation (3), if we do not consider regularization.

However, due to the dependency of a conditional density estimate on a training data set, the conditional density estimation approach often suffers from non-negligible epistemic (i.e. model) uncertainty and bias in the construction of the conditional density estimates $p(z|\mathbf{f})$. This can lead to sub-optimal probability calibration of the estimates $p(z|\mathbf{f})$. Appendix A describes an estimating function approach that allows the marginalization over the epistemic (or 'model uncertainty') and aleatoric (or 'intrinsic statistical noise') uncertainty in the estimator construction of equation (4). This is achieved via the formulation of a likelihood function.

5 PHOTOMETRIC REDSHIFT INFERENCE PIPELINE

In the following subsections we describe in more detail our methodology for performing $p_{\rm samp}(z)$ inference for HSC Y3 WL analyses. We reiterate that all estimates for $p_{\rm samp}(z)$ in this work include the lensing weights that are available for all galaxies in the shear catalogue as described in Section 3.1.

5.1 Individual Galaxy redshift estimation

In the following text, we will briefly describe the three photometric redshift techniques for individual galaxies used in this work. For a more detailed description of these methods we refer to the photometric redshift analysis study for the third public data release.[9]

5.1.1 Mizuki

The photometric redshift code Mizuki (Tanaka 2015; Tanaka et al. 2018) is an SED fitting technique. It uses an SED template set constructed using Bruzual–Charlot models (Bruzual & Charlot 2003), a stellar population synthesis code that uses an initial mass function following Chabrier (2003), a dust attenuation modelling from Calzetti et al. (2000), and emission-line modelling assuming solar metallicity (Inoue 2011). The method applies a set of redshift-dependent Bayesian priors on the physical properties. After estimation, the photometric redshift distributions of galaxies are calibrated (Bordoloi, Lilly & Amara 2010) using the specXphot data set to improve error quantification. We refer the reader to

[8]Note that the likelihood is not a probability density, but a function. The probability measure is 'provided' by the prior.

[9]https://hsc-release.mtk.nao.ac.jp/doc/wp-content/uploads/2022/08/pdr3_photoz.pdf (Accessed 2022 October)

Tanaka (2015) and Tanaka et al. (2018) for more details on the methodology.

5.1.2 DNNz

DNNz is a neural-network-based photometric redshift conditional density estimation code. The DNNz architecture consists of multi-layer perceptrons with five hidden layers. The training uses cmodel fluxes, unblended convolved fluxes, point spread function fluxes, and galaxy shape information. The construction of the conditional density uses 100 nodes in the output layer, and each represent a redshift histogram bin spanning from z = 0 to 7 (Nishizawa et al. in preparation).

5.1.3 DEMPz

The Direct Empirical Photometric redshift code (DEMPz) is an empirical technique for photometric redshift estimation (Hsieh & Yee 2014; Tanaka et al. 2018) that constructs conditional density estimates. The technique uses quadratic polynomial interpolation of 40 nearest neighbour galaxies in a training set, with a distance estimated in a 10-dimensional feature space (5 mag, four colours, and shape information). DEMPz obtains error estimates for the constructed conditional densities using resampling procedures. This also includes resampling of the input feature uncertainties and bootstrapping the training galaxies.

5.2 Sample selection

We bin the full sample described in Section 3.1 into four tomographic bins by selecting galaxies using the best estimation of the DNNz conditional density estimates within redshift intervals of (0.3, 0.6], (0.6, 0.9], (0.9, 1.2], and (1.2, 1.5].

After catalogue creation we identify regions of data space that will be difficult to calibrate using the cross-correlations with the CAMIRA LRG sample, and therefore have the potential to produce a large systematic error (see Section 5.5). In particular, we identify double solutions in the Mizuki SED fits and DNNz conditional density estimates, associated with a significant fraction of outliers at $z \gtrsim 3.0$ for both methods. These photometric redshift solutions have redshift-template degeneracies that produce multiple solutions. Since the secondary solutions are outside the redshift coverage of the CAMIRA LRG sample, they cannot be calibrated using spatial cross-correlations. Therefore, we decide to remove these galaxies from the sample.

We identify galaxies with double solutions by defining the following selection metric based on the distance between the 0.025 and 0.975 quantiles of the Mizuki posterior solutions and DNNz conditional density estimates:

$$\left(z^{\rm Mizuki}_{0.975,i} - z^{\rm Mizuki}_{0.025,i}\right) < 2.7 \quad \text{and} \quad \left(z^{\rm DNNz}_{0.975,i} - z^{\rm DNNz}_{0.025,i}\right) < 2.7\,, \qquad (5)$$

where $z^{\rm Mizuki}_{0.975,i}$ and $z^{\rm Mizuki}_{0.025,i}$ denote the 0.975 and 0.025 percentiles for galaxy i derived using the Mizuki estimates of posterior redshift, respectively; and similarly for the DNNz conditional density redshift predictions. We found that the above criteria based on the Mizuki and DNNz methods is optimal to ensure that the removal of double solutions is efficient for Mizuki, DNNz, and DEMPz.

We apply this criterion to the first and the second tomographic redshift bins, reducing their sample size by 31 per cent and 8 per cent, respectively. The third and fourth tomographic bins have negligible double solutions. We therefore do not apply any cuts to the corresponding galaxy samples. We illustrate the effect of removing the double solutions on the stacked (summed) redshift distribution in Fig. 6. We can see that a reduction of 31 per cent in sample size by applying equation (5) removes double solutions for all three methods available in this work. In the following text, we will denote the removal of double solutions as the 'calibration cut'.

We have also confirmed that this selection does not induce a spatial selection effect. This was tested by comparing the spatial distribution of galaxies before and after we apply the calibration cut and confirming that no significant modification of the clustering was introduced by the cut.

This is illustrated in Fig. 7, where we test the impact of the calibration cut on the spatial distribution of galaxies. We first confirm if the fraction of galaxies rejected by the calibration cut (i.e. galaxies with doubly peaked $p_{\rm indiv}(z)$) s is comparable for all subfields. This has to take into account the variation due to sampling variance, which we quantify by dividing into subregions within the different fields. The top panel plots several normalized histograms over s where each histogram corresponds to a separate field listed in the legend. Note that we obtain a distribution $p(s)$ over s for each field by estimating s on each patch within each field. The vertical dashed line denotes the mean of the histograms over the different fields, the errorbars denote the field-to-field variation. We see that s is consistent across the different fields.

In the lower panels we investigate if the spatial distribution of removed galaxies is spatially 'random', or if we have to expect a correlation signal based on the calibration cut. The vertical axis shows the difference between the correlation function estimated on the catalogue in each field subject to the calibration cut and a catalogue where galaxies are removed randomly. The horizontal dashed line guides the eye towards the zero line. The error contours are obtained by jackknife resampling the catalogue within each field. We see that the measured autocorrelation functions are consistent between the randomly selected catalogue and the catalogue subject to the calibration cut.

5.3 Individual Galaxy redshift estimation to enable sample redshift distribution ($p_{\rm samp}(z)$) inference

This project considered all three individual galaxy photometric redshift estimates introduced in Section 5.1 and performed an initial comparison between sample redshift posteriors obtained using these three methods with the cross-correlation constraints. We found insufficient agreement for the Mizuki solutions, whereas DEMPz and DNNz where more consistent. By iteratively reproducing the inconsistencies using analytic error models, we identified a number of problems with the Mizuki solutions.

We found that the Mizuki photometric redshift solutions are miscalibrated (Nishizawa et al. in preparation) and that systematics induced by uncorrected selection functions from galaxy selection, object weighting, and the calibration cut can lead to additional bias in the sample redshift inference for the Mizuki code. A recalibration of the Mizuki likelihoods using the specXphot sample based on Bordoloi et al. (2010) only slightly improved the results. We concluded that the consistency between the DEMPz and DNNz codes and the cross-correlation measurements was still better. We note that including the aforementioned selection function into the likelihood formulation is structurally simple, but would require a rerun of the Mizuki solutions which was not deemed practical. We

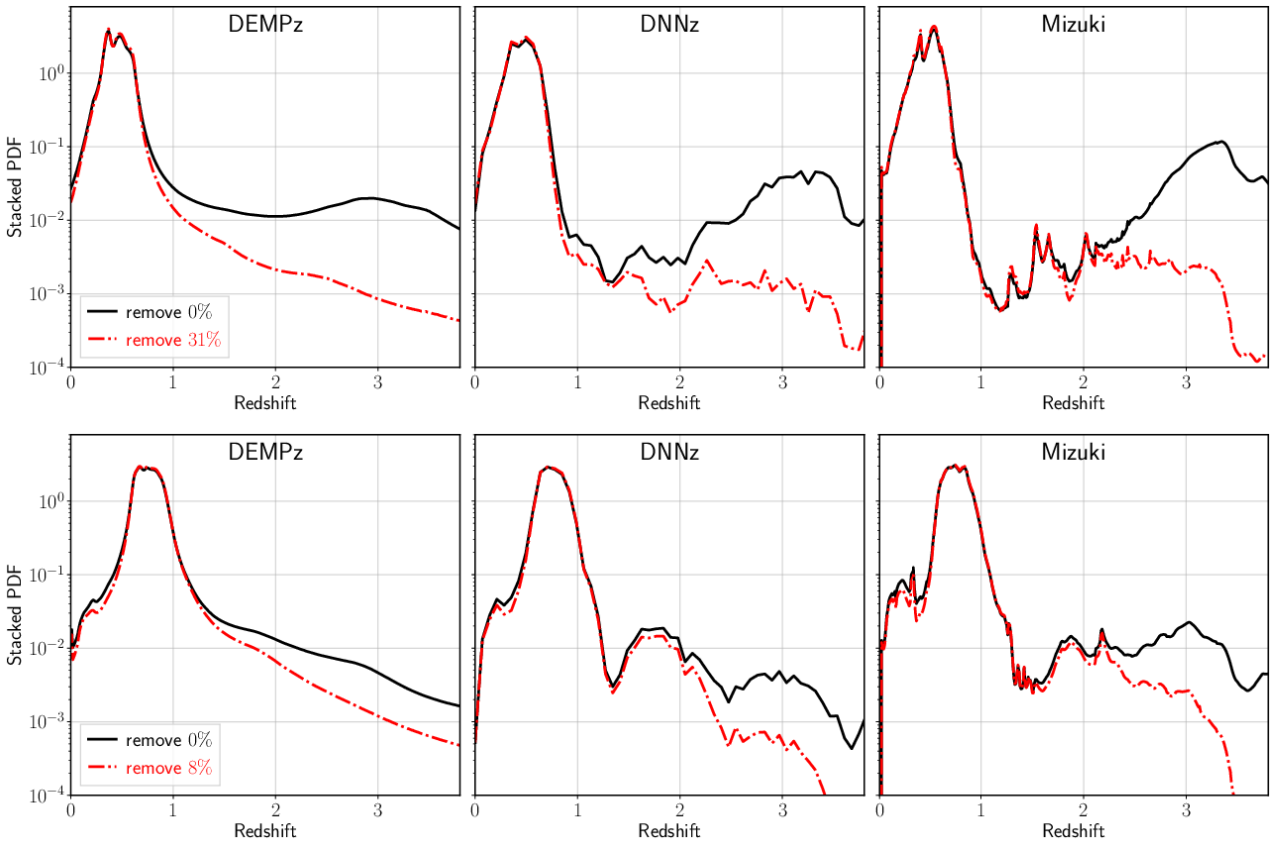

Figure 6. The stacked photo-z posteriors for galaxies in the first (upper panel, $0.3 < z_{\rm dnnz\_best} \leq 0.6$) and second (lower panel, $0.6 < z_{\rm dnnz\_best} \leq 0.9$) tomographic redshift bin estimated from three photo-z estimation codes. Cuts on the interquartile distance are applied to these galaxies to remove the secondary peak in the stacked posteriors. The stacked posteriors for the fiducial cut, which removes 31 per cent of the galaxies in the first bin, are plotted as red lines. These posteriors are normalized so that they have total probability of one.

therefore selected DNNz as our primary method and DEMPz as the alternative method for the subsequent analysis. In the following text, we will refer to sample redshift distribution inference methodology based on individual galaxy redshift distributions, abbreviated as the vector-valued $\vec{p}_{\rm indiv}(z)$, as 'photometry-based $p_{\rm samp}(z)$ estimation', or short 'PhotZ'.

5.4 Formulation of the ensemble redshift distribution prior

Based on our fiducial model choice we apply the empirical likelihood methodology described in Appendix A to estimate $p_{\rm samp}(z)$ for the four tomographic bins based on the DNNz $\vec{p}_{\rm indiv}(z)$.

As we discuss in detail in Appendix A, the empirical likelihood estimation obeys the central limit theorem. The large sample size of our catalogues implies that the statistical error in the maximum empirical likelihood estimate is much smaller than other sources of uncertainty. These include a cosmic variance contribution from the spatially limited training sample (see Section 3.3), as well as the uncertainty in the individual galaxy redshift estimation model (epistemic uncertainty). In the remainder of this section we will discuss our approach to including cosmic variance into our sample redshift estimation procedure. Our treatment of the epistemic uncertainty will be discussed in Section 5.7.

The basis for our $p_{\rm samp}(z)$ error model is the logistic Gaussian process. The logistic Gaussian process, first applied to sample redshift estimation by Rau et al. (2020), assumes that the number counts of galaxies as a function of redshift are lognormally distributed. The model can capture cross-bin correlations and provides more modelling complexity than, e.g. the Dirichlet distribution as we discuss in Appendix E.

The logistic Gaussian process prior on the parameters $\boldsymbol{\phi}_{\rm nz}$ can be formulated as follows:

$$\begin{aligned} \boldsymbol{s} &\sim N(\boldsymbol{s}|\mu, \Sigma) \\ \boldsymbol{\rho} &= \exp(\boldsymbol{s}) \\ \boldsymbol{\phi}_{\rm nz} &:= \left\{ \frac{\rho_i}{\sum_j \rho_j} \,\middle|\, 0 < i < N_{\rm bins} \right\}, \end{aligned} \tag{6}$$

where $(\boldsymbol{\mu}/\boldsymbol{\Sigma})$ denotes the (mean vector/covariance matrix). We note that equation (6) relates to a lognormal model for the galaxy counts, where ρ is the expected number of galaxies per redshift. The dimension of $(\boldsymbol{s}/\boldsymbol{\rho}/\boldsymbol{\phi}_{\rm nz})$ is $N_{\rm bins}$ as introduced in equation (1).

As discussed in Section 3.3, the faint end of our training set is dominated by COSMOS2015 data. This induces a cosmic variance error contribution that we include into our logistic Gaussian process model based on the cosmic variance measurements for the COSMOS2015 data set by Sánchez et al. (2020). We detail our methodology in Appendix C.

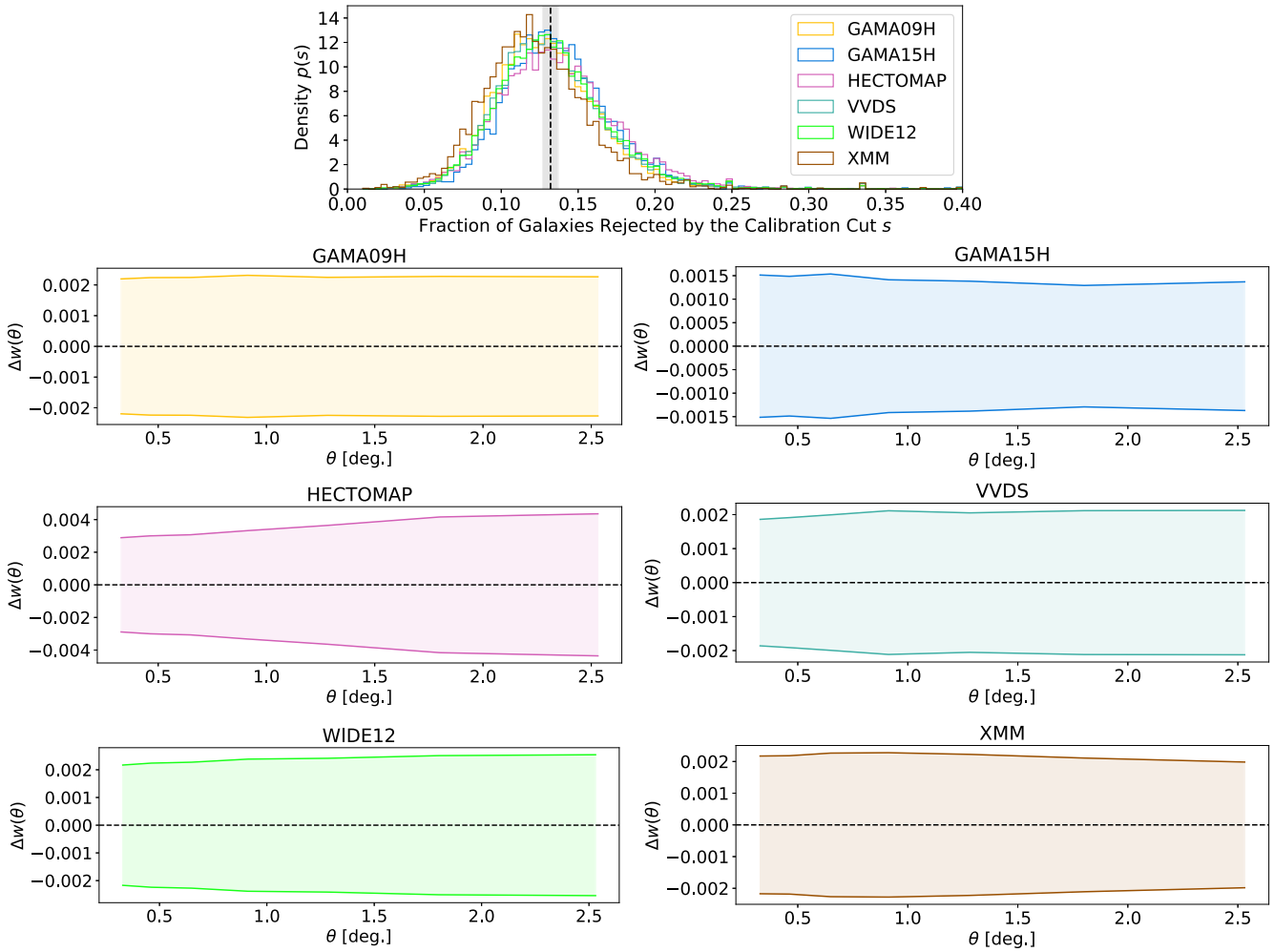

Figure 7. Testing the impact of the calibration cut on the spatial distribution of galaxies by resampling the catalogue for the first tomographic bin. Top panel: Test if the fraction of galaxies rejected by the calibration cut (s) is comparable for all subfields. Each histogram corresponds to a separate field listed in the legend, where the histograms over s show the variation across the different patches within the field. The vertical dashed line denotes the mean of the histograms over the different fields with errorbars denoting the field-to-field variation. Lower panels: Testing if the spatial distribution of removed galaxies is 'random'. The vertical axis shows the difference between the correlation function estimated on the catalogue in each field subject to the calibration cut and a catalogue where galaxies are removed randomly. The horizontal dashed line shows the zero line. The error contours are obtained by jackknife resampling.

5.5 Ensemble redshift distribution likelihood from spatial cross-correlations (cross-correlation)

To further constrain the $p_{\rm samp}(z)$, we utilize spatial cross-correlations with the CAMIRA LRG sample. This approach has two goals: it provides an independent consistency check for the $p_{\rm samp}(z)$ derived using the DNNz approach, and it allows a joint inference of the $p_{\rm samp}(z)$ informed by both the photometry of galaxies and the spatial cross-correlations with the CAMIRA LRG sample.

As detailed in Section 3.2, the CAMIRA LRG sample extends only to $z \lesssim 1.2$ and the photoZ of the CAMIRA LRG galaxies are themselves subject to error. This subsection gives an overview of the cross-correlation measurements and the likelihood formulation. We refer to Appendix D for the technical details.

Using vector notation, where each vector component corresponds to the cross-correlation measurement in a redshift bin, we can predict the spatial cross-correlation between the CAMIRA LRG sample and HSC phot as

$$\mathbf{w}_{\rm LRG-Y3} \propto \boldsymbol{\phi}_{\rm nz}\, \mathbf{b}_{\rm PhotZ}\, \mathbf{b}_{\rm LRG}\, \mathbf{w}_{\rm DM}\,, \tag{7}$$

where $\mathbf{w}_{\rm DM}$ is the scale-averaged, redshift- and cosmology-dependent, two-point function of the dark matter density field. The terms $\mathbf{b}_{\rm PhotZ}$ and $\mathbf{b}_{\rm LRG}$ are the redshift-dependent galaxy-dark matter bias terms from the (HSC phot/CAMIRA LRG) sample and $\boldsymbol{\phi}_{\rm nz}$ are the parameters defined in equation (1). We use 'The-Wizz' (a code described in Morrison et al. 2017) to measure these cross-correlations and use bootstrap re-sampling (as described in Morrison et al. 2017) to obtain a covariance matrix of the measurements. We include the lensing weights in the two-point estimator, and choose a scale range of 0.1–1.0 Mpc for our measurements. These measurements are repeated for 10 catalogues generated by sampling from our CAMIRA LRG photometric error model, which is a conditional density estimate that maps the noisy CAMIRA LRG photometric redshift to the unknown true redshifts. This mapping is trained on the specXphot calibration data.

Using the scheme described in Appendix D we marginalize over the realizations to derive a likelihood for the cross-correlation measurements that has an inflated covariance $\boldsymbol{\Sigma}_{\rm LRG-PhotZ}$ due to the contribution of the CAMIRA LRG photometric redshift error. Using

a Gaussian Likelihood ansatz we obtain

$$p(\hat{\mathbf{w}}_{\mathrm{LRG-PhotZ}}|\boldsymbol{\phi}_{\mathrm{nz\,PhotZ}}, \mathbf{b}_{\mathrm{PhotZ}}, \mathbf{b}_{\mathrm{LRG}}) = N(\hat{\mathbf{w}}_{\mathrm{LRG-PhotZ}}|\mathbf{w}_{\mathrm{LRG-PhotZ}}(\boldsymbol{\phi}_{\mathrm{nz\,PhotZ}}, \mathbf{b}_{\mathrm{PhotZ}}, \mathbf{b}_{\mathrm{LRG}}), \boldsymbol{\Sigma}_{\mathrm{LRG-PhotZ}}), \tag{8}$$

where $\hat{\mathbf{w}}_{\mathrm{LRG-PhotZ}}$ denotes the spatial cross-correlation measurements between the CAMIRA LRG and HSC phot catalogues, $\mathbf{w}_{\mathrm{LRG-PhotZ}}(\boldsymbol{\phi}_{\mathrm{nz\,PhotZ}}, \mathbf{b}_{\mathrm{PhotZ}}, \mathbf{b}_{\mathrm{LRG}})$ denotes the theory prediction, and $\boldsymbol{\Sigma}_{\mathrm{LRG-PhotZ}}$ the covariance matrix that is adjusted for the CAMIRA LRG photometric redshift error.

In this analysis we marginalize over a parameter that describes the product $\mathbf{b}_{\mathrm{PhotZ}}\,\mathbf{b}_{\mathrm{LRG}}$ for each tomographic bin. For three tomographic bins we therefore have three parameters that account for the product of galaxy-dark matter bias for galaxies in the HSC phot and the CAMIRA LRG samples. We predict[10] the dark matter contribution $\mathbf{w}_{\mathbf{DM}}$ using the *Core Cosmology Library, version 1.0.0* (CCL, Chisari et al. 2019)[11] using halofit to model the non-linear power spectrum (Takahashi et al. 2012). We do not marginalize over cosmological parameters that enter $\mathbf{w}_{\mathrm{LRG\text{-}PhotZ}}$, as we find that the choice of cosmology does not strongly impact the posterior $p_{\mathrm{samp}}(z)$. Concretely, we note that the spatial cross-correlation data vector is a scale-averaged correlation function. Its redshift scaling affects the inferred cross-correlation redshift distributions on the ~ 20 per cent level by (suppressing/increasing) the (low/high)-z flank. However, variations in cosmology affect the redshift scaling of the scale-averaged dark-matter correlation at the ~ 10 per cent level (for rather extreme cosmologies at the 2σ contour of Stage III surveys), which implies that the cosmology-dependence of the inferred cross-correlation redshift distributions is subdominant to other systematics such as the redshift-dependent galaxy-dark matter bias modelling uncertainties.

5.6 Joint constraints

Using the logistic Gaussian Process model defined in Section 5.4 and the cross-correlation likelihood defined in equation (8), we can sample from the joint posterior of the parameters that describe the sample redshift distribution $\boldsymbol{\phi}_{\mathrm{nz}}$, defined in equation (1), and the product $\boldsymbol{b} = \boldsymbol{b}_{\mathrm{LRG}}\,\boldsymbol{b}_{\mathrm{PhotZ}}$ of the galaxy-dark matter bias of the CAMIRA LRG ($\boldsymbol{b}_{\mathrm{LRG}}$) and HSC phot ($\boldsymbol{b}_{\mathrm{PhotZ}}$) samples

$$p(\boldsymbol{\phi}_{\mathrm{nz}}, \boldsymbol{b}|\hat{\mathbf{w}}_{\mathrm{LRG-PhotZ}}) \propto p(\hat{\mathbf{w}}_{\mathrm{LRG-PhotZ}}|\boldsymbol{\phi}_{\mathrm{nz}}, \boldsymbol{b})p(\boldsymbol{\phi}_{\mathrm{nz}})p(\boldsymbol{b})\,. \tag{9}$$

The sampling of the $\boldsymbol{\phi}_{\mathrm{nz}}$ parameters has to be carried out with respect to a likelihood that only constrains a subset of $\boldsymbol{\phi}_{\mathrm{nz}}$ due to the limited redshift coverage of the CAMIRA LRG sample. We note that the parameters $\boldsymbol{\phi}_{\mathrm{nz}}$ can be normalized to lie on the simplex[12], i.e. to sum to unity. It is therefore useful to instead perform inference with respect to the random variable $\mathbf{s}$, defined in equation (6). Using this reparametrization we can perform inference in $\mathbb{R}^{\mathrm{N_{bins}}}$ using standard approaches and then transform to the original parameter $\boldsymbol{\phi}_{\mathrm{nz}}$. We use Elliptical Slice Sampling (Murray, Adams & MacKay 2010) for our inference. Elliptical slice sampling works particularly well for a logistic Gaussian process prior, since it can utilize the aforementioned reparametrization that relates the logistic Gaussian process to the multivariate normal distribution.

Fig. 8 shows the resulting posterior sample redshift distributions for the following three scenarios:

(i) photometry-based sample redshift distribution estimation ('PhotZ (DNNz)', grey) utilizing the DNNz code and including our model for cosmic variance following Section 5.3 and Section 5.4;
(ii) clustering redshift estimation ('WX (0.1–1.0 Mpc)', black) following Section 5.5; and
(iii) the combination of spatial information and photometry ('PhotZ & WX', red) following Section 5.6.

The horizontal axis of Fig. 8 shows the redshift, while the vertical axis shows the probability density of posterior tomographic $p_{\mathrm{samp}}(z)$. The distributions are normalized to integrate to unity. We report contours/errorbars corresponding to piecewise $\pm 1\sigma$ errors. In the case of 'PhotZ' and 'PhotZ & WX' which both have asymmetric posterior distributions, we report contours between the 16th and 84th percentiles. The blue errorbars show the standard deviation in the mean[13] cross-correlation measurement with respect to the different catalogue draws from the CAMIRA LRG error model. We specifically see that even for only 10 catalogues, this error is already much smaller compared with the statistical uncertainty of 'WX (0.1–1.0 Mpc)'. We note that the black errorbars for the cross-correlation constraints are plotted assuming the maximum a posteriori values of $\boldsymbol{b}$ defined in equation (9), which act to normalize the clustering redshift measurements. This allows us to plot the clustering redshift constraints on the same scale as 'PhotZ (DNNz)' and 'PhotZ & WX'. We note that we do this for illustrative purposes only; we marginalize over $\boldsymbol{b}$ to infer 'PhotZ & WX'.

Since the CAMIRA LRG sample redshift coverage extends to $z < 1.2$, we can only partially calibrate the third tomographic bin. We also decided to not include a cross-correlation data vector in the sample redshift distribution calibration of the fourth tomographic bin. This is motivated by the overall small redshift overlap with the CAMIRA LRG sample. Furthermore, for significant parts of the relevant redshift range ($1.0 < z < 1.2$), there is a trend in the inferred $n(z)$ in the third bin that might indicate the need for more complex modelling of astrophysical effects like redshift-dependent galaxy-dark matter bias. It is therefore likely that we might include additional systematics in the calibration of the fourth tomographic bin low-redshift tail for very moderate gains in statistical accuracy.

We conclude that the clustering redshift measurements are broadly consistent with the constraints we derive based on the photometry of galaxies. However, there are slight inconsistencies between the 'PhotZ' and 'WX' constraints around $z \approx 0.2$. This is around the same redshift where we know that the photometric redshift distributions of the CAMIRA LRG galaxies are biased (see Section 3.2). This implies an incomplete correction of this bias from our error model. This inconsistency is moderate, on the level of 2σ–3σ with respect to the joint posterior (PhotZ & WX). We leave further investigations for future work.

5.7 Prior recommendation for WL analysis

Fig. 9 shows the distribution of posterior mean for the four tomographic bins. We define the posterior mean as the mean estimated for each posterior tomographic $p_{\mathrm{samp}}(z)$ sample. We can derive the distribution of posterior mean for each tomographic bin by sampling from the posterior $p_{\mathrm{samp}}(z)$ shown in Fig. 9. This is done for the joint

[10] We use $\Omega_{\mathrm{DM}} = 0.258868$, $\Omega_{\mathrm{b}} = 0.048252$, $h = 0.6777$, $n_s = 0.95$, and $\sigma_8 = 0.8$.
[11] https://github.com/LSSTDESC/CCL (Accessed 2022 September 22)
[12] The probability simplex is defined as $S = \{x_i | \sum_{i=1}^{N} x_i = 1 \text{ and } 0 \leq x_i \leq 1 \text{ for } 1 \leq i \leq N\}$.
[13] We refer here to the standard deviation in the mean estimate, which scales with $1/\sqrt{N}$, where N corresponds to the number of catalogues drawn.

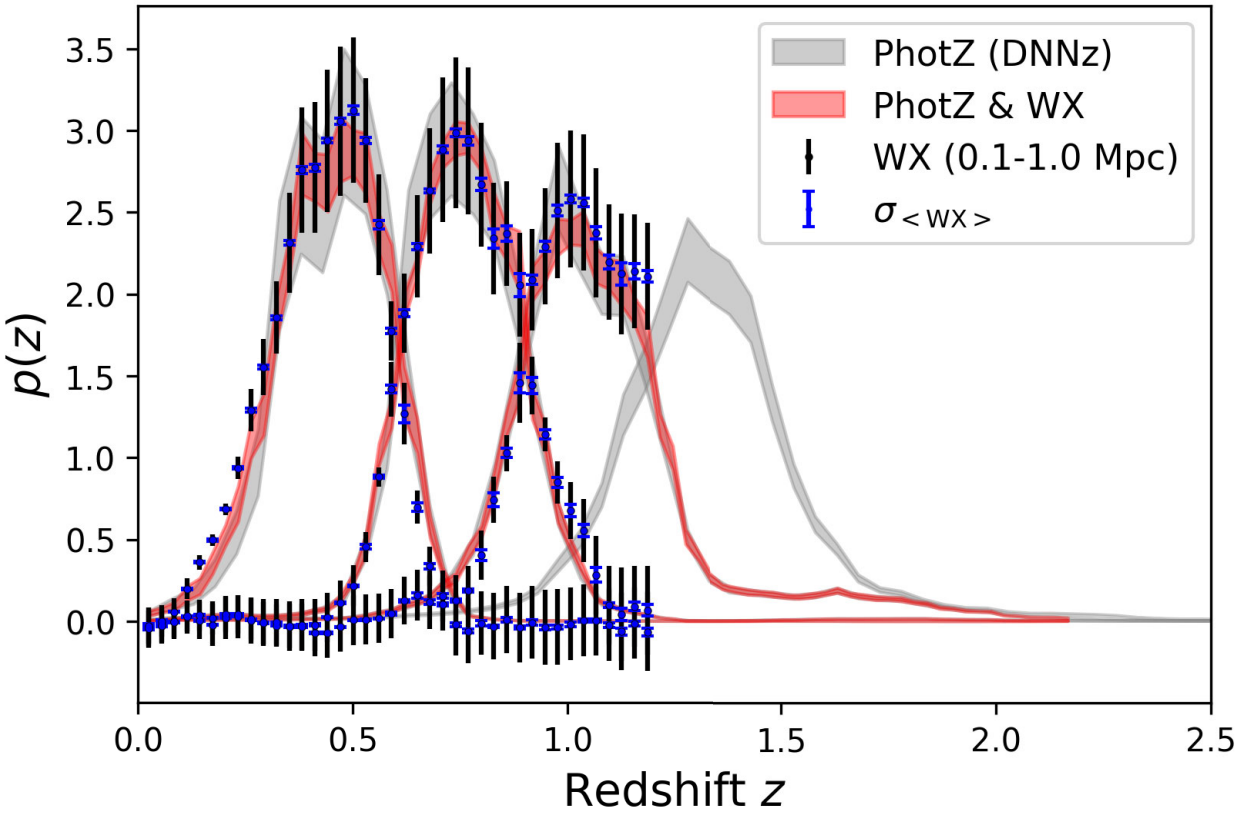

Figure 8. Sample redshift distribution ($p_{\rm samp}(z)$) posteriors for the four tomographic redshift bins of the HSC Y3 lensing sample derived to include information from the photometry ('PhotZ (DNNz)', grey area), spatial clustering ('WX (0.1–1.0 Mpc)', black dots) and the combination of spatial information and photometry ('PhotZ & WX', red area). The blue dots denote the standard deviation on the mean of WX (i.e. clustering redshift) measurements. The CAMIRA LRG sample has a limited redshift coverage to $z < 1.2$, due to which the high-redshift tomographic bin does not include a cross-correlation data vector. The inference includes the lensing weights consistently in all likelihood terms. The piecewise intervals denote the $\pm 1\sigma$ errors.

constraint ('PhotZ & WX', red contours) and the photometry-based inference ('PhotZ (DNNz)', grey contours) for each tomographic bin. We now estimate the mean of each sample drawn in this way. This results in distributions of posterior mean for our tomographic bins in both scenarios.

We see that the distributions of posterior mean are consistent for the two methods in the first three tomographic bins. There is mild tension in the lowest tomographic bin, which can be explained by the inconsistency at $z \approx 0.2$ as described in the previous section. We further quantify the 'information gained' by the cross-correlation likelihood over the 'PhotZ (DNNz)' prior by calculating the Kullback–Leibler (KL) Divergence between the prior and posterior based on the results quoted in Table 1, where we use a Gaussian approximation for the posterior mean distributions of tomographic bins to calculate the KL Divergence. The KL divergence between prior and posterior is referred to as the 'Bayesian Surprise' in statistics (see e.g. Itti & Baldi 2009; Baldi & Itti 2010)[14] and the results are quoted in Table 1 under the column 'Bayesian Surprise'. Table 1 indicates that the largest amount of information is added in the first tomographic bin. We note, however, that this does not allow us to judge if the Bayesian Surprise is due to unaccounted systematics or statistical fluctuation. A comparison with the results from the second and third bins, which are an order of magnitude smaller, hints towards unaccounted systematics in the first bin as the most likely explanation for the large Bayesian surprise value.

Fig. 9 further illustrates that the width of the distributions of posterior mean decreases when we include the spatial cross-correlation data vector. This highlights the importance of including cross-correlations in the sample redshift calibration as both a consistency check and an additional constraint. We relate this result to the expected biases in the WL power spectra in Fig. 10, proceeding in close analogy to our study of the distribution of posterior mean. We estimate the WL power spectra on each draw from the posterior $p_{\rm samp}(z)$ using the *Core Cosmology Library, version 1.0.0* (CCL, Chisari et al. 2019)[15] and calculate the relative bias ΔC_ℓ between the posterior distributions of WL power spectra estimated using the photometry alone (Phot (DNNz)) and including the spatial cross-correlations (Phot & WX). The relative bias is defined as

$$\Delta C_\ell = \frac{C_\ell^{\rm Phot\,\&\,WX} - C_\ell^{\rm Phot\,(DNNz)}}{C_\ell^{\rm Phot\,(DNNz)}} \,. \tag{10}$$

Fig. 10 shows ΔC_ℓ as a function of scale for the (first/second/third) tomographic bin. We see that the relative difference between the measurements using the photometry (DNNz) alone shows a tension

[14]The Bayesian Surprise is sometimes referred to as the 'information gain' in cosmology (e.g. Grandis et al. 2016).

[15]https://github.com/LSSTDESC/CCL (Accessed 2022 September 22)

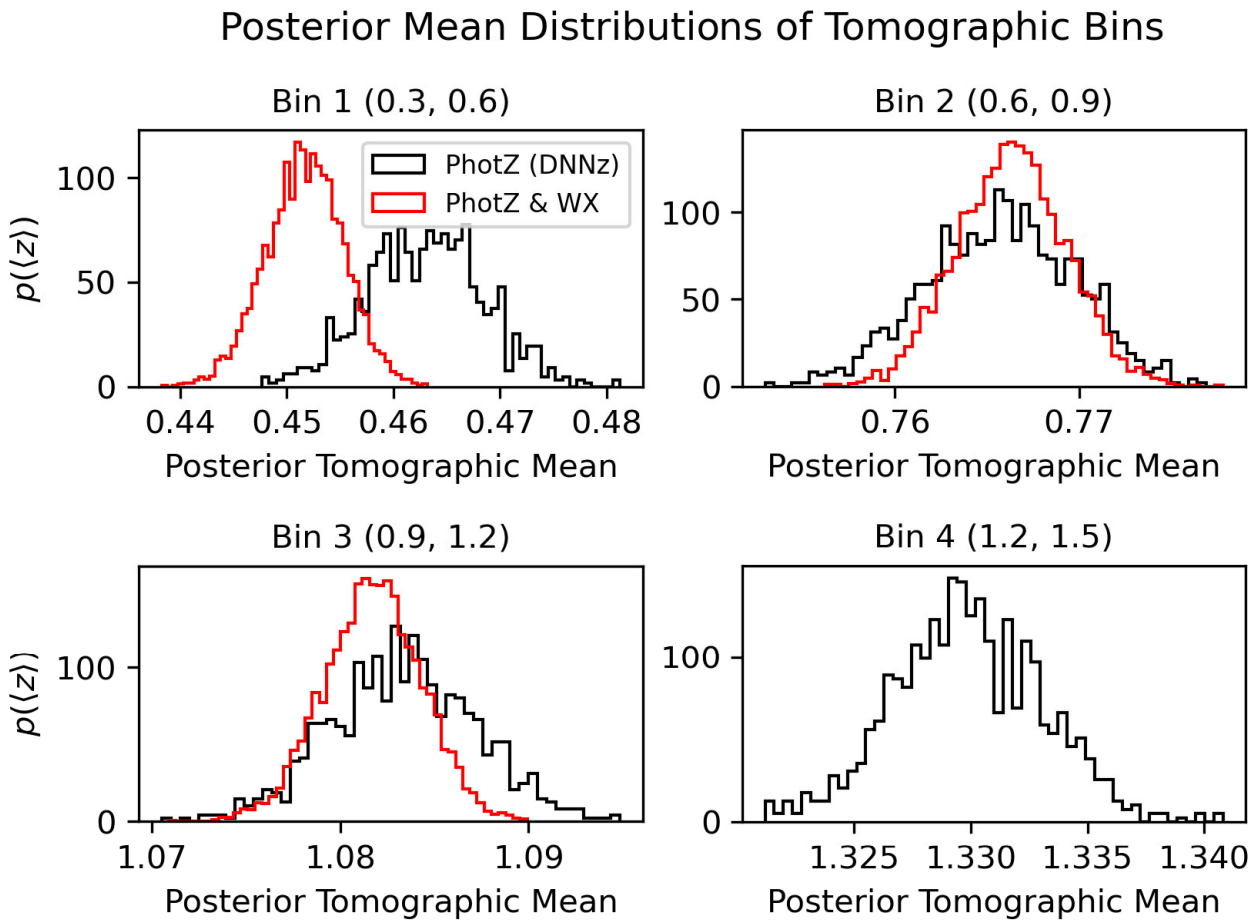

Figure 9. Comparison of the distributions of posterior tomographic mean for the four tomographic redshift distributions shown in Fig. 9. The subpanels correspond to increasing tomographic bin mean redshift. The (red/black) histograms show the result for the constraint (PhotZ (DNNz)/Phot & WX) which corresponds to the exclusion/inclusion of the spatial cross-correlation data vector with the CAMIRA LRG sample. There is consistency between the posterior distributions of tomographic mean estimates obtained using the photometry alone (black) and in combination with the clustering redshift data vector (red).

Table 1. Mean and standard deviation of the posterior mean for the different tomographic redshift bins. The first column lists the corresponding results for the first year analysis (Hamana et al. 2020) (Y1 Analysis), the results obtained using the photometry alone with cosmic variance correction (PhotZ (DNNz)), the results we obtain using the DEMPz code (Y3 DEMPz), and the joint constraints with the cross-correlation data vector (Y3 PhotZ & WX). The DEMPz results, here used as an alternative methodology, are obtained by taking the average of the normalized $\vec{p}_{\rm indiv}(z)$. For conditional density estimates like DEMPz this amounts to a mean estimate of the marginalization in equation (4) (see Section 4.2). The final two columns lists the Bayesian Surprise values (Y3 Bayesian Surprise) and the total error budget that includes our systematics error budget as explained in Section 5.7 (Y3 Total). We note that all columns except the first are derived on the year 3 data set described in Section 3 with different galaxy selection (but similar redshift range) compared with the S16A analysis.

	Y1 Analysis	Y3 PhotZ (DNNz)	Y3 DEMPz	Y3 PhotZ & WX	Y3 Bayesian Surprise	Y3 Total
Bin 1	0.44 (0.0285)	0.463 (0.005)	0.463	0.452 (0.004)	3.84	0.452 (0.024)
Bin 2	0.77 (0.014)	0.766 (0.004)	0.777	0.766 (0.003)	0.10	0.766 (0.022)
Bin 3	1.05 (0.0383)	1.084 (0.004)	1.097	1.081 (0.004)	0.28	1.081 (0.031)
Bin 4	1.33 (0.0376)	1.330 (0.003)	1.350	-	-	1.330 (0.034)

that is significant in Bin 1 compared with the expected signal-to-noise ratio, which hints towards remaining uncorrected systematic biases. We discuss this further in Section 7. In the following text we discuss our conservative assessment of tomographic $p_{\rm samp}(z)$ error motivated by the aforementioned tensions.

Since the sample redshift posteriors obtained in this work will be used as part of the HSC Y3 WL cosmological analysis, we discuss here which parametrization we will employ to marginalize over sample redshift uncertainty. Following Zhang et al. (2023), we will use the maximum a posteriori solution for the $p_{\rm samp}(z)$ and vary the mean using a Gaussian prior informed by the inference described in the previous sections. While Zhang et al. (2023) explored multiple ways of marginalizing over the full posterior for the redshift distribution, at the level of precision of this HSC analysis, marginal-

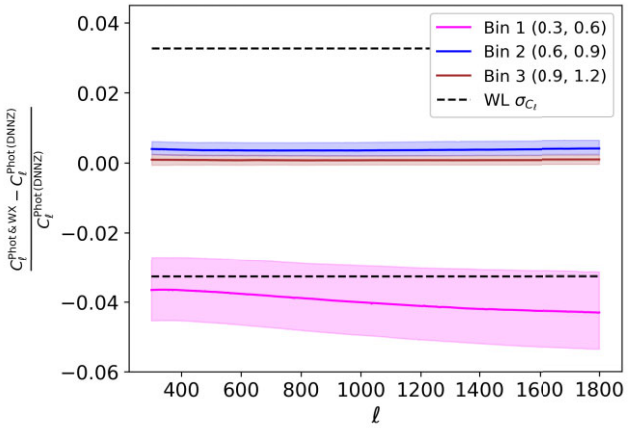

Figure 10. Comparison of the distributions of relative bias in WL power spectra (see equation 10) between the posterior $p_{\text{samp}}(z)$ informed by the photometry alone (Phot (DNNz)) and the joint constraints that include the spatial cross-correlations (Phot & WX). We plot the results for the (first/second/third) tomographic bins (Bin 1/Bin 2/Bin 3) corresponding to the results shown in Figs 8 and 9. The solid lines show the median and the contours show the (16/84) percentiles corresponding to the Gaussianized $\pm 1\sigma$ errors. The black horizontal dashed lines show the $\pm 1\sigma$ errors that correspond to the expected signal-to-noise ratio of the WL power spectra measurements.

izing over uncertainty in the mean redshift was found to be entirely sufficient. We also include an additional error contribution that parametrizes differences in sample redshift inference across different $\vec{p}_{\text{indiv}}(z)$ solutions, where we will use DEMPz as an alternative method.

We derive the combined error budget based on the aforementioned parametrization of the posterior mean. In order to include discrepancies between different $\vec{p}_{\text{indiv}}(z)$ solutions into the analysis, we compare the results obtained using DNNz with the DEMPz results. The DEMPz method was selected because it showed superior photometric redshift accuracy compared with the Mizuki results[16] and overall better consistency with the clustering redshift measurements.

Since the DEMPz and DNNz methods will be correlated, we have to formulate an upper limit on the error budget. Furthermore, we require that this upper limit calculation will be conservative with respect to the residual systematics in Bin 1 discussed in Fig. 9 and the HSC first-year (Y1) result (Hamana et al. 2020) for Bin 4, as Bin 4 lacks the additional constraints from the spatial correlations with the CAMIRA LRG sample.

While we present a significantly updated methodology, we do not provide additional data-driven consistency checks that would warrant a significantly smaller systematic error budget compared with the Y1 analysis. To derive this total error budget we combine the standard deviation of the posteriors of the joint constraint (shown as red histograms in Fig. 9), which we will denote as σ_{joint}, with the absolute difference between the $p_{\text{samp}}(z)$ derived using the alternative method DEMPz and our joint fiducial analysis. The latter error contribution will be denoted as σ_{sys}. We reiterate that we consider here only the posterior tomographic mean.

[16]See https://hsc-release.mtk.nao.ac.jp/doc/wp-content/uploads/2022/08/pdr3_photoz.pdf

We introduce the correlation coefficient ρ with $|\rho| \leq 1$ and combine σ_{sys} with the statistical error budget σ_{joint} as

$$\begin{aligned}\sigma_{\text{joint,sys}} &= \sqrt{\sigma_{\text{joint}}^2 + \sigma_{\text{sys}}^2 + 2\rho\sigma_{\text{sys}}\sigma_{\text{joint}}} \\ &\leq \sqrt{\sigma_{\text{joint}}^2 + \sigma_{\text{sys}}^2 + 2\sigma_{\text{sys}}\sigma_{\text{joint}}}\,. \end{aligned} \tag{11}$$

An upper limit on $\sigma_{\text{joint, sys}}$ is therefore given as $\sigma_{\text{joint, sys}} \leq \sigma_{\text{joint}} + \sigma_{\text{sys}}$. This implies an upper limit for (Bin 1/Bin 4) of $\sigma_{\text{joint, sys, (Bin1/Bin4)}} = (0.015/0.023)$. This systematic error budget for the Bin 1 and Bin 4 is similar to the absolute difference between the constraints of 'PhotZ (DNNz)' and 'Phot & WX' in Fig. 9 and much smaller than the error budget for Bin 4 assumed in Y1 as quoted in Table 1. We therefore choose to utilize a more conservative upper limit by applying the Minkowsi inequality directly to equation (11):

$$\sigma_{\text{joint,sys}} \leq \sigma_{\text{joint}} + \sigma_{\text{sys}} + \sqrt{2\sigma_{\text{sys}}\sigma_{\text{joint}}}\,. \tag{12}$$

We recommend the right-hand side of equation (12) as a conservative prior width for the HSC Y3 cosmological WL analysis. However, we strongly recommend performing a sensitivity study for this prior width especially for Bin 4. We refer to Dalal et al. (2023), Li et al. (2023a), More et al. (2023), Miyatake et al. (2023), and Sugiyama et al. (2023) for further details on the conclusions of this analysis and their implications on prior choices.

Table 1 summarizes our results by giving the mean and standard deviation of the posterior mean for the various analysis scenarios presented in this work. The columns list the corresponding results for the Y1 analysis in Hamana et al. (2020), the results obtained for HSC Y3 using the photometry alone with cosmic variance correction ('PhotZ (DNNz)' in Fig. 8), the results we obtain using the DEMPz code and the joint constraints that include the cross-correlation data vector ('PhotZ & WX' in Fig. 8). The final column lists the final result that includes the conservative assessment of model error following equation (12).

The error budget we obtain from a combination of cross-correlations and photometry without the additional systematic uncertainty term is almost an order of magnitude smaller than in the HSC Y1 results. The $p_{\text{samp}}(z)$ constraints we obtain from the cross-correlation measurements and the $\vec{p}_{\text{indiv}}(z)$ are consistent. The model error assessment that we use for our final recommendation on priors is therefore very conservative and is very similar and/or more conservative compared with the error budget in the HSC Y1 analysis. We note that the error budget is dominated by our assessment of model error, i.e. derived by the comparison with the DEMPz method. This assessment of model error is conservative, since the joint constraint between the CAMIRA LRG and the photometry based inference would allow for almost an order of magnitude smaller error in the posterior mean.

However, it is not overly pessimistic and is less than double the residual systematic expected from the difference between the PhotZ (DNNz) and PhotZ&WX results presented in Bin 1 of Fig. 9. Future work would benefit from adding additional constraints to the high-redshift tomographic bin, e.g. by including spatial cross-correlations with DESI and by reconsidering the low-redshift systematics in the cross-correlation constraints.

6 SUMMARY

This work presents posterior sample redshift distributions ($p_{\text{samp}}(z)$) in four tomographic bins for the HSC three-year shape catalogue. To exploit the synergy between complementary sources of redshift information, we combined $p_{\text{samp}}(z)$ constraints from spatial cross-

correlations and from individual galaxy photometric redshift distributions ($\vec{p}_{\rm indiv}(z)$) derived from the galaxies photometry. We perform cross-correlation based $p_{\rm samp}(z)$ inference using the CAMIRA LRG sample, which allowed us to obtain constraints within the limited redshift range of the LRG sample of $z \leq 1.2$. The presented analysis had to account for three main sources of systematic biases and uncertainties: the intrinsic photometric redshift error in the LRGs, the significant variation (both methodologically and in quality) of the provided $\vec{p}_{\rm indiv}(z)$, and the spatial colour-redshift-dependent selection functions of our specXphot redshift calibration sample.

The goals of the analysis were to provide posteriors for the relevant tomographic $p_{\rm samp}(z)$, demonstrate consistency between the constraints derived using the spatial cross-correlations and $\vec{p}_{\rm indiv}(z)$, and recommend priors on $p_{\rm samp}(z)$ parameters for the cosmological WL analysis. The latter should also incorporate an assessment of model error and should reflect conservative analysis choices under acceptable degradation of cosmological parameter constraints. We claim that these analysis goals were accomplished in our analysis.

Our analysis was structured as follows (see Section 5):

(i) Sample definition and selection (Section 5.2);
(ii) Estimation of individual and tomographic $p_{\rm samp}(z)$ using photometry-based inference (Phot, Section 5.3);
(iii) Incorporation of cosmic variance from the spatially limited specXphot training sample into the constraint (Section 5.4);
(iv) Cross-correlation-based $p_{\rm samp}(z)$ inference (WX, Section 5.5);
(v) Joint inference combining WX and Phot (Section 5.6);
(vi) Recommendation of the science-ready photometric redshift priors for WL (Section 5.7).

The sample was limited to galaxies with single-peaked $\vec{p}_{\rm indiv}(z)$. The removal of galaxies that show secondary, high-redshift ($z > 1.2$) photometric redshift solutions is essential for our analysis, to ensure that we can validate our photometric redshifts with the data products available. Since the CAMIRA LRG sample does not allow a calibration to $z > 1.2$ and the specXphot calibration sample is expected to be incomplete at the faint end of the colour–magnitude space, we cannot reliably validate secondary solutions at $z \gtrapprox 1.2$.

This work introduces a framework for sample redshift inference for both empirical methods based on conditional density estimation and methods that are based on SED fitting or likelihood-based forward modelling. Initially we considered three methods for $\vec{p}_{\rm indiv}(z)$ estimation: a likelihood based SED fitting code (Mizuki) and two empirical methods (DNNz, DEMPz).

We selected the DNNz method, a conditional density estimation method for photometric redshifts, as our fiducial inference method based on initial comparisons with the cross-correlation data vector. As the specXphot calibration sample used for training the individual galaxy redshift estimators at the faint end of the sample covers only a small solid angle, we construct a logistic Gaussian Process model to parametrize the cosmic variance component in the error model for the inferred tomographic $p_{\rm samp}(z)$.

In the next analysis step, we measured spatial cross-correlations between the CAMIRA LRG and the HSC Y3 photometric shape catalogue (HSC phot) for the first three tomographic bins (within $z < 1.2$) and account for the photometric redshift error in the CAMIRA LRG sample in the construction of the cross-correlation likelihood. We demonstrated consistency between the $p_{\rm samp}(z)$ constraints derived from the cross-correlation data vector and photometry-based sample redshift inference.

Utilizing a joint inference framework that accounts for the limited redshift coverage of the cross-correlation measurements, we obtained posterior $p_{\rm samp}(z)$ in four tomographic bins.

Finally we included a conservative error assessment based on a comparison with an alternative photometric redshift algorithm, 'DEMPz'. While the final constraint on the mean of the tomographic bins is much narrower than the results obtained in the HSC Y1 analysis (Hamana et al. 2020), our conservative assessment of model error yields a prior recommendation for the HSC three-year WL analysis that is similar to (and more conservative than) the Y1 HSC cosmological WLanalysis.

7 DISCUSSION AND FUTURE WORK

In the following text, we describe a range of known limitations in our analysis that motivate our conservative error assessment and highlight avenues for future work. We concentrate on five areas of this analysis where we identified limitations:

(i) Error quantification of $\vec{p}_{\rm indiv}(z)$;
(ii) Treatment of selection functions of the specXphot calibration sample;
(iii) Treatment of cosmic variance induced by redshift calibration using the specXphot calibration sample;
(iv) Photometric redshift uncertainties and systematics of CAMIRA LRG galaxies;
(v) Simplistic treatment of astrophysical effects in the modelling of the cross-correlation data vector.

In the following paragraphs we will discuss each of these items in order.

(i) There are a number of unmodelled systematics in the construction of $\vec{p}_{\rm indiv}(z)$ using DNNz, DEMPz, and Mizuki that are likely explanations for the large differences between their estimates relative to the statistical uncertainty. We show this in Table 1 where the model error from differences in the DNNz and DEMPz results dominates the error budget. This is qualitatively consistent with the first year HSC analysis[17] of individual galaxy redshift distribution systematics in Tanaka et al. (2018). Figs 11 and 14 in that paper illustrate significant differences between the estimates obtained using different methodologies both in terms of the estimated $p_{\rm samp}(z)$ (fig. 11) and in terms of the PIT metric (fig. 14), which quantifies how well the $\vec{p}_{\rm indiv}(z)$ are calibrated with respect to a specXphot reference data set. The significant differences between the methods imply an incomplete assessment of model error.[18]

(ii) While the specXphot calibration data were assembled to reduce the impact of unwanted selection functions and we employ the calibration cut (see Section 5.2) to remove problematic regions in colour space with doubly peaked $p_{\rm indiv}(z)$, it likely does not provide an unbiased source of redshift calibration for model evaluation and training. Our analysis therefore used cross-correlations with the CAMIRA LRG sample, within the aforementioned limited redshift coverage, for redshift calibration and imposed a conservative assessment of model error. The latter is motivated by an acceptable degradation in the cosmological parameter constraints forecasted for the upcoming WL analysis. However, future analyses with the full HSC survey data set and upcoming surveys such as LSST will have

[17] Tanaka et al. (2018) analyse Y1 data. The paper does not present a principled inference strategy to derive $p_{\rm samp}(z)$, e.g. Mizuki that requires deconvolving for photometric redshift error (see Section 4.1). However, this does not invalidate a qualitative comparison with our analysis.
[18] Model error refers here to error contributions (both systematic and statistical), for example from lack of training data, uncorrected selection functions in the training data, inaccurate modelling of SEDs, priors, or photometry.

to continue to further improve the analysis methodology to reduce this source of systematic uncertainty.

(iii) Our approach to quantify cosmic variance from the spatially small calibration field suffers from three main limitations that we discuss in the following text. We note, however, that the current analysis will likely not be methodologically limited in this area as the dominant source of uncertainty is the model error in the $\vec{p}_{\rm indiv}(z)$. The modelling of the variance of the point field within a patch on the sky depends not only on the point-field expected number density per area and redshift, which can be scaled to match the colour-redshift distribution of the target field, but also on the clustering of the galaxies of the underlying process. The latter is modelled based on the COSMOS2015 field, which covers a small area, has different clustering properties than other fields, and might be subject to a non-random spatial selection function.[19] The small area and nonrandom selection function implies that any statistic derived from this field will not be fully representative of other fields. This means that our cosmic variance estimate derived on COSMOS2015 is not necessarily representative of the true cosmic variance contribution of photometric redshift estimates trained on any COSMOS2015-size patch on the sky. Since the galaxy field is ergodic, this becomes less of a concern for spatially larger fields or if several small but spatially separated fields are used. Furthermore, since the variance does not uniquely identify the stochastic process that describes the $p_{\rm samp}(z)$ uncertainty, every assessment of cosmic variance has model assumptions. We discuss this point in detail in Appendix E. We note that we neglect spatial correlations between the COSMOS2015 field and HSC phot, i.e. we do not formulate a full spatial model for redshift inference in this work, which can affect our assessment of cosmic variance. These limitations affect the redshift calibration in other surveys such as DES, which is also based on spatially small calibration fields. We also note that the individual galaxy redshift estimates presented in this work do not allow us to construct a direct relation to the COSMOS2015 training set galaxies, which limits our ability to perform a cosmic variance correction in colour space. In future work, we will present a spatial model for redshift inference that will extend the current approach to treat cosmic variance in $p_{\rm samp}(z)$ estimation (Rau et al. in preparation).

(iv) Our modelling of the WX data vector depends on accurately parametrizing the photometric redshift systematics of the CAMIRA LRG sample. As discussed, especially at low redshift, these systematics can be quite significant. Our current modelling is based on a specXphot calibration sample, as we did not obtain access to the relevant CAMIRA LRG likelihoods. As a result, our correction could be subject to residual systematics from spectroscopic selection functions. This needs to be reconsidered in the future, along with a better assessment of galaxy-dark matter bias for the calibration sample. This includes parametrizing a redshift and scale dependence in the galaxy-dark matter bias within each tomographic bin for the photometric sample and the calibration sample. In order to constrain this more complex assessment of galaxy-dark matter bias, it will be important to extend the data vector towards autocorrelations of the photometric and reference samples.

(v) Regarding the modelling of the cross-correlation data vector, we limited our analysis to a constant galaxy-dark matter bias within each tomographic source bin and did not include an assessment of magnification bias. Gatti et al. (2022) studied the effect of magnification bias on cross-correlation based $p_{\rm samp}(z)$ inference in the context of the Dark-Energy-Survey Year 3 analysis. While performed in the context of different data and analysis, we can expect the effect of magnification bias to be subdominant compared with the modelling of a redshift-dependent galaxy-dark matter bias and subdominant compared with our conservative total error budget. While based on a qualitative extrapolation of their quantitative assessments (see Table Gatti et al. 2022), the good agreement between WX and Y3 PhotZ reported in Fig. 8 provide some basis for that claim. Future measurements with larger signal-to-noise ratio will need to reconsider this assumption.

In conclusion, we have presented a $p_{\rm samp}(z)$ inference methodology for the HSC Y3 shape catalogue that represents a significant update over the methodology in previous HSC WL analyses. We have forecasted the effect of our updated methodology on the previous HSC S16A analysis in Section 2 and demonstrated that our updated methodology can account for shifts in the $\Omega_{\rm m}$-S_8 plane of 0.5σ after rescaling the covariance matrix from previous HSC WL measurements to account for the increased area in the HSC Y3 catalogue. This highlights the importance of sample redshift calibration as we prepare not only for the HSC analysis but also look ahead towards upcoming surveys like LSST.

ACKNOWLEDGEMENTS

We thank the anonymous referee for helpful comments that improved both content and presentation of the paper. The HSC Collaboration acknowledges fundamental work on photometric redshifts by the Complete Calibration of the Color Redshift Relation (C3R2) team. The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from the Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University.

This paper uses software developed for Vera C. Rubin Observatory. We thank the Rubin Observatory for making their code available as a free software at http://pipelines.lsst.io/.

This paper is based on data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by the Subaru Telescope and Astronomy Data Center (ADC) at NAOJ. Data analysis was in part carried out with the cooperation of Center for Computational Astrophysics (CfCA), NAOJ. We are honored and grateful for the opportunity of observing the Universe from Maunakea, which has the cultural, historical, and natural significance in Hawaii.

The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg, and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edin-

[19]For example, randomly selecting multiple spatially small patches on the sky would show different clustering properties than 'favouring interesting' regions with an abundance of clusters and therefore produce a different cosmic-variance model.

burgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

Work at Argonne National Laboratory was supported by the U.S. Department of Energy, Office of High Energy Physics. Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated by UChicago Argonne LLC under contract no. DE-AC02-06CH11357. MMR acknowledges the Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. MMR's work at Argonne National Laboratory was also supported under the U.S. Department of Energy contract DE-AC02-06CH11357. RD acknowledges support from the NSF Graduate Research Fellowship Program under Grant No. DGE-2039656. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. AJN is supported by Grant-in-Aid for Transformative Research Areas 21H05454, and JSPS KAKENHI Grant Numbers JP20H0193 and JP21K03625. RM is supported by DOE grant DE-SC0010118 and a grant from the Simons Foundation (Simons Investigator in Astrophysics, Award ID 620789). MT is supported by World Premier International Research Center Initiative (WPI Initiative), JSPS KAKENHI Grant Numbers JP20H05850, JP20H05855, and JP19H00677 and by Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

DATA AVAILABILITY

This work is part of the HSC Year 3 cosmological analysis. The data and analysis products, as well as the software, will be made publicly available via the HSC-SSP website https://hsc.mtk.nao.ac.jp/ssp/data-release/.

REFERENCES

Abazajian K. et al., 2004, AJ, 128, 502
Abbott T. M. C. et al., 2018, ApJS, 239, 18
Abbott T. M. C. et al., 2022, Phys. Rev. D, 105, 023520
Aihara H. et al., 2018, PASJ, 70, S4
Aihara H. et al., 2022, PASJ, 74, 247
Alam S. et al., 2015, ApJS, 219, 12
Alarcon A., Sánchez C., Bernstein G. M., Gaztañaga E., 2020, MNRAS, 498, 2614
Albrecht A. et al., 2006, preprint(astro–ph/0609591)
Amon A. et al., 2022, Phys. Rev. D, 105, 023514
Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, 540
Asgari M. et al., 2021, A&A, 645, A104
Baldi P., Itti L., 2010, Neural Netw., 23, 649
Benítez N., 2000, ApJ, 536, 571
Benjamin J. et al., 2013, MNRAS, 431, 1547
Bernstein G., Huterer D., 2010, MNRAS, 401, 1399
Bonnett C., 2015, MNRAS, 449, 1043
Bordoloi R., Lilly S. J., Amara A., 2010, MNRAS, 406, 881
Bradshaw E. J. et al., 2013, MNRAS, 433, 194
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Calzetti D., Armus L., Bohlin R. C., Kinney A. L., Koornneef J., Storchi-Bergmann T., 2000, ApJ, 533, 682
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Cawthon R. et al., 2022, MNRAS, 513, 5517
Chabrier G., 2003, PASP, 115, 763
Chang C. et al., 2016, MNRAS, 459, 3203
Chisari N. E. et al., 2019, ApJS, 242, 2
Clerkin L., Kirk D., Lahav O., Abdalla F. B., Gaztañaga E., 2015, MNRAS, 448, 1389
Coil A. L., Weiner B. J., Holz D. E., Cooper M. C., Yan R., Aird J., 2011, ApJ, 743, 46
Collister A. A., Lahav O., 2004, PASP, 116, 345
Cool R. J. et al., 2013, ApJ, 767, 118
Dalal R. et al., 2023, preprint (arXiv:2304.00701)
Dalmasso N., Pospisil T., Lee A. B., Izbicki R., Freeman P. E., Malz A. I., 2020, Astron. Comput., 30, 100362
Davis M. et al., 2003, in Guhathakurta P.ed., Proc. SPIE Conf. Ser. Vol. 4834, Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II. SPIE, Bellingham, p. 161
Davis C. et al., 2017, preprint (arXiv:1710.02517)
Drinkwater M. J. et al., 2010, MNRAS, 401, 1429
Feldmann R. et al., 2006, MNRAS, 372, 565
Garilli B. et al., 2014, A&A, 562, A23
Gatti M. et al., 2018, MNRAS, 477, 1664
Gatti M. et al., 2022, MNRAS, 510, 1223
Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823
Giblin B. et al., 2021, A&A, 645, A105
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Grandis S., Seehars S., Refregier A., Amara A., Nicola A., 2016, J. Cosmol. Astropart. Phys., 2016, 034
Greisel N., Seitz S., Drory N., Bender R., Saglia R. P., Snigula J., 2015, MNRAS, 451, 1848
Hamana T. et al., 2020, PASJ, 72, 16
Hartley W. G. et al., 2020, MNRAS, 496, 4769
Heymans C. et al., 2021, A&A, 646, A140
Hikage C. et al., 2019, PASJ, 71, 43
Hildebrandt H. et al., 2017, MNRAS, 465, 1454
Hildebrandt H. et al., 2021, A&A, 647, A124
Hoyle B., 2016, Astron. Comput., 16, 34
Hoyle B., Rau M. M., 2019, MNRAS, 485, 3642
Hoyle B. et al., 2018, MNRAS, 478, 592
Hsieh B. C., Yee H. K. C., 2014, ApJ, 792, 102
Huang S. et al., 2017, PASJ, 70
Huterer D., Takada M., Bernstein G., Jain B., 2006, MNRAS, 366, 101
Huterer D., Lin H., Busha M. T., Wechsler R. H., Cunha C. E., 2014, MNRAS, 444, 129
Ilbert O. et al., 2006, A&A, 457, 841
Inoue A. K., 2011, MNRAS , 415, 2920
Ishikawa S., Okumura T., Oguri M., Lin S.-C., 2021, ApJ, 922, 23
Itti L., Baldi P., 2009, Vis. Res., 49, 1295
Ivezić Ž. et al., 2019, ApJ, 873, 111
Joachimi B. et al., 2021, A&A, 646, A129
Jones D. M., Heavens A. F., 2019, MNRAS, 483, 2487
Joudaki S. et al., 2020, A&A, 638, L1
Kohonen T., 1982, Biol. Cybern., 43, 59
Laigle C. et al., 2016, ApJS, 224, 24
Laureijs R. et al., 2011, preprint (arXiv:1110.3193)
Le Fèvre O. et al., 2013, A&A, 559, A14
Leistedt B., Mortlock D. J., Peiris H. V., 2016, MNRAS, 460, 4258
Li X. et al., 2022, PASJ, 74, 421
Li X. et al., 2023a, preprint (arXiv:2304.00702)
Li S.-S. et al., 2023b, A&A, 670, A100
Lilly S. J. et al., 2009, ApJS, 184, 218
Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118

Liske J. et al., 2015, MNRAS, 452, 2087
Lupton R., Gunn J. E., Ivezić Z., Knapp G. R., Kent S., 2001, in Harnden F. R. J., Primini F. A., Payne H. E.eds, ASP Conf. Ser. Vol. 238, Active Galaxies. Astronomical Data Analysis Software and Systems X. Astron. Soc. Pac., San Francisco, p. 269
Ma Z., Hu W., Huterer D., 2006, ApJ, 636, 21
MacCrann N. et al., 2022, MNRAS, 509, 3371
Malz A. I., Hogg D. W., 2020, preprint (arXiv:2007.12178)
Mandelbaum R., 2018, ARA&A, 56, 393
Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes J. D., Castander F. J., Paltani S., 2017, ApJ, 841, 111
Masters D. C. et al., 2019, ApJ, 877, 81
Matarrese S., Coles P., Lucchin F., Moscardini L., 1997, MNRAS, 286, 115
McLeod M., Balan S. T., Abdalla F. B., 2017, MNRAS, 466, 3558
McLure R. J. et al., 2013, MNRAS, 428, 1088
McQuinn M., White M., 2013, MNRAS, 433, 2857
Meister A., 2009, Deconvolution Problems in Nonparametric Statistics Lecture Notes in Statistics. Springer, Berlin Heidelberg, available at: https://link.springer.com/book/10.1007/978-3-540-87557-4
Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, preprint (arXiv:1303.4722)
Minka T. P., 2000, Estimating a Dirichlet Distribution
Miyatake H. et al., 2023, preprint (arXiv:2304.00704)
Miyazaki S. et al., 2018, PASJ, 70, S1
Momcheva I. G. et al., 2016, ApJS, 225, 27
More S. et al., 2023, preprint (arXiv:2304.00703)
Morrison C. B., Hildebrandt H., Schmidt S. J., Baldry I. K., Bilicki M., Choi A., Erben T., Schneider P., 2017, MNRAS, 467, 3576
Murray I., Adams R., MacKay D., 2010, in Teh Y. W., Titterington M.eds, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics Vol. 9. PMLR, Chia Laguna Resort, Sardinia, Italy, p. 541, available at: https://proceedings.mlr.press/v9/murray10a.html
Myles J. et al., 2021, MNRAS, 505, 4249
Newman J. A., 2008, ApJ, 684, 88
Newman J. A., Gruen D., 2022, ARA&A, 60, 363
Newman J. A. et al., 2013, ApJS, 208, 5
Newman J. A. et al., 2015, Astropart. Phys., 63, 81
Nishizawa A. J., Hsieh B.-C., Tanaka M., Takata T., 2020, preprint (arXiv:2003.01511)
Oguri M., 2014, MNRAS, 444, 147
Oguri M. et al., 2018a, PASJ, 70, S20
Oguri M. et al., 2018b, PASJ, 70, S26
Owen A., 1990, Ann. Stat., 18, 90
Owen A., 2001, Empirical Likelihood Chapman and Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, Florida, available at: https://books.google.com/books?id = tHbLBQAAQBAJ
Padmanabhan N. et al., 2005, MNRAS, 359, 237
Pandey S. et al., 2022, Phys. Rev. D, 106, 043520
Pawitan Y., 2001, In All Likelihood: Statistical Modelling and Inference Using Likelihood. OUP Oxford, available at: https://books.google.com/books?id = M-3pSCVxV5oC
Pentericci L. et al., 2018, A&A, 616, A174
Prat J. et al., 2018, MNRAS, 473, 1667
Prat J. et al., 2019, MNRAS, 487, 1363
Prat J. et al., 2022, Phys. Rev. D, 105, 083528
Raccanelli A., Rahman M., Kovetz E. D., 2017, MNRAS, 468, 3650
Rau M. M., Seitz S., Brimioulle F., Frank E., Friedrich O., Gruen D., Hoyle B., 2015, MNRAS, 452, 3710
Rau M. M., Hoyle B., Paech K., Seitz S., 2017, MNRAS, 466, 2927
Rau M. M., Wilson S., Mandelbaum R., 2020, MNRAS, 491, 4768
Rau M. M., Morrison C. B., Schmidt S. J., Wilson S., Mandelbaum R., Mao Y. Y., Mao Y. Y., LSST Dark Energy Science Collaboration, 2022, MNRAS, 509, 4886
Rosenblatt M., 1956, Ann. Math. Stat., 27, 832
Salvato M., Ilbert O., Hoyle B., 2019, Nat. Astron., 3, 212
Sánchez C., Bernstein G. M., 2019, MNRAS, 483, 2801
Sánchez C., Raveri M., Alarcon A., Bernstein G. M., 2020, MNRAS, 498, 2984–2999
Sánchez C. et al., 2022, Phys. Rev. D, 105, 083529
Scottez V. et al., 2016, MNRAS, 462, 1683
Scranton R. et al., 2005, ApJ, 633, 589
Secco L. F. et al., 2022a, Phys. Rev. D, 105, 023515
Secco L. F. et al., 2022b, Phys. Rev. D, 105, 023515
Silverman J. D. et al., 2015, ApJS, 220, 12
Simon P., Hilbert S., 2018, A&A, 613, A15
Skelton R. E. et al., 2014, ApJS, 214, 24
Spergel D. et al., 2015, preprint (arXiv:1503.03757)
Stölzner B., Joachimi B., Korn A., Hildebrandt H., Wright A. H., 2021, A&A, 650, A148
Stoyan D., Stoyan H., 1994, Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics. Wiley, Hoboken, available at: https://books.google.com/books?id = Dw3vAAAAMAAJ
Stölzner B., Joachimi B., Korn A., the LSST Dark Energy Science Collaboration, 2022, MNRAS , 519, 2438
Sugiyama S., Takada M., Kobayashi Y., Miyatake H., Shirasaki M., Nishimichi T., Park Y., 2020, Phys. Rev. D, 102, 083520
Sugiyama S. et al., 2023, preprint (arXiv:2304.00705)
Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C., Giordano G., 2003, Neural Networks for Photometric Redshifts Evaluation, Springer Publishing , New York, p. 226
Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152
Tanaka M., 2015, ApJ, 801, 20
Tanaka M. et al., 2018, PASJ, 70, S9
van den Busch J. L. et al., 2020, A&A, 642, A200
Zhang T., Rau M. M., Mandelbaum R., Li X., Moews B., 2023, MNRAS, 518, 709

APPENDIX A: CONDITIONAL DENSITY ESTIMATION METHODOLOGY

A1 Overview

In the following appendix, we describe our methodology to perform sample redshift inference in the context of conditional density estimation in continuation of Section 4.2. The discussion in this section applies to the DNNz and DEMPz methods. The basic idea of conditional density redshift estimation is to construct an estimator of the true conditional density $p(z|\mathbf{f})$ of the redshift z given the fluxes (or photometry) $\mathbf{f}$. We construct this mapping between the 'true' redshift z and measured flux, which requires a specXphot 'training' data set. This can be constructed using spatially overlapping spectroscopic and photometric survey data, which provides both photometry and accurate spectroscopic redshifts. Upon constructing a conditional density estimator $\hat{p}(z|\mathbf{f})$, for a particular photometric survey, we can construct an estimator of the $p_{\rm samp}(z)$ as

$$\hat{p}_{\rm samp}(z) = \int \hat{p}(z|\mathbf{f})\hat{p}(\mathbf{f})\mathrm{d}\mathbf{f}\,, \tag{A1}$$

where $\hat{p}(z|\mathbf{f})$ and $\hat{p}(\mathbf{f})$ denote estimators of the conditional density of redshift z given flux $\mathbf{f}$ and of the marginal density of colour–magnitude space $\hat{p}(\mathbf{f})$. We note here the difference between constructing an estimator of the conditional density $\hat{p}(z|\mathbf{f})$ and a 'forward modelling' approach that would require the formulation of a likelihood (or the non-parametric estimator thereof) $p(\mathbf{f}|z)$. The former is a density estimation problem and requires the availability of a calibration data set to provide information on the redshift distribution of galaxies as a function of measured photometry. The latter induces an inverse problem that depends on knowledge of the data-generating process from a true redshift to measured photometry. One would include additional redshift information here in the formulation of the prior. We stress that these model formulations are very different and refer

to Appendix B for a detailed description of the redshift inference methodology in the context of likelihood-based forward modelling. In the following text, we will comment on the assumptions behind the conditional density estimation methodology.

A2 Assumptions

The basic assumption of empirical methods is that the data in the calibration and target data sets follow the same conditional densities $p(z|\mathbf{f})$ of the redshift z given the fluxes (or photometry) $\mathbf{f}$. We also note that there exist no unbiased non-parametric density estimators (Rosenblatt 1956). Therefore, a prime challenge for these methods is the selection of bandwidth, or smoothing scale.

Inaccurate selection of smoothing can lead to biases in redshift inference that are relevant for cosmological inference, as shown in prior work (Rau et al. 2017). The authors also demonstrated that biases from inaccurate selection of smoothing can be mitigated in cosmological inference using parametric bootstrap techniques. In the context of this work, we can assume that misspecification errors due to non-representative training data and epistemic uncertainty from a lack of training data will be more severe than biases due to inaccurate bandwidth selection.

A3 Methodology

In the following text, we construct an empirical likelihood of the density estimator equation (A1) that allows us to marginalize over systematics in a principled way. Under the assumptions described in the previous paragraph, we can parametrize $p(z|\mathbf{f})$ as a histogram

$$p(z|\mathbf{f}) = \sum_{i=1}^{N_{\rm bins}} w_i(\mathbf{f}) \mathbb{1}_i(z)\,, \tag{A2}$$

where w_i denotes the histogram bin height and $\mathbb{1}_i(z)$ is unity if the redshift is within bin i, and zero otherwise. $N_{\rm bins}$ denotes the number of histogram bins.

This yields an estimator for $p(z)$ as

$$p(z) = \sum_{i=1}^{N_{\rm bins}} \left(\int w_i(\mathbf{f}) p(\mathbf{f}) {\rm d}\mathbf{f} \right) \mathbb{1}_i(z) = \sum_{i=1}^{N_{\rm bins}} E_{\mathbf{f}}\left[w_i(\mathbf{f})\right] \mathbb{1}_i(z)\,, \tag{A3}$$

where $E_{\mathbf{f}}[w_i(\mathbf{f})]$ denotes the expectation value of the weights $w_i(\mathbf{f})$ wrt to the marginal distribution of photometry. The weights $w_i(\mathbf{f})$ can depend on parameters $\boldsymbol{\eta}$ that describe additional sources of error, induced by unmodelled selection functions in the training data or by intrinsic model bias in the conditional density estimates.

Based on this relation, we can employ the empirical likelihood formalism (e.g. Owen 1990, 2001; Pawitan 2001) and construct an estimating equation

$$\boldsymbol{\psi}([\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\eta}], \mathbf{f}) = \boldsymbol{w}(\mathbf{f}, \boldsymbol{\eta}) - \boldsymbol{\phi}_{\rm nz}, \tag{A4}$$

where $\boldsymbol{\phi}_{\rm nz}$ denotes the modelled histogram heights (see equation 1) of the $p_{\rm samp}(z)$. We note here that $\boldsymbol{w}(\mathbf{f}, \boldsymbol{\eta})$ is a function of the measured photometry and parameters that describe other systematics, whereas $\boldsymbol{\phi}_{\rm nz}$ is the parameter vector to be estimated.

Under the assumption that the parameter set $\boldsymbol{\eta}$ accurately describes the systematics mentioned above, we seek values for $\boldsymbol{\eta}$ and $\boldsymbol{\phi}_{\rm nz}$ such that

$$E_{\mathbf{f}}\left[\boldsymbol{\psi}([\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\eta}], \mathbf{f})\right] = 0\,. \tag{A5}$$

We can treat the application of lensing weights $\omega_{\rm lens}(\mathbf{f})$ as a selection function and follow the recipe described in Owen (2001) of modifying the expected estimating equation by transforming the probability measure as

$$0 = \int \boldsymbol{\psi}([\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\eta}], \mathbf{f}) dF(\mathbf{f}) = \int \boldsymbol{\psi}_{\rm WL}([\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\eta}], \mathbf{f}) \omega_{\rm lens}(\mathbf{f}) dG(\mathbf{f})\,, \tag{A6}$$

where $\omega_{\rm lens}(\mathbf{f})$ denotes the lensing weights as a function of photometry (and other auxillary parameters omitted here). In the following text, we will omit the dependence of the lensing weights on $\mathbf{f}$ for convenience. It is understood that the introduction of lensing weights implies a dependence on a variety of parameters that describe the measurement of galaxy shapes.

$({\rm d}F(\mathbf{f})/{\rm d}G(\mathbf{f}))$ denotes the (unweighted/weighted) probability measures where ${\rm d}F(\mathbf{f}) = \omega_{\rm lens}{\rm d}G(\mathbf{f})$.

We introduce $(\boldsymbol{w}_{\mathbf{WL}}(\mathbf{f}, \boldsymbol{\eta})/\boldsymbol{\phi}_{\rm nz\,WL})$ that denote the weighted (measured/modelled) WL histogram height parameters that include lensing weights as

$$\begin{aligned} \boldsymbol{w}_{\mathbf{WL}}(\mathbf{f}, \boldsymbol{\eta}) &= \boldsymbol{w}(\mathbf{f}, \boldsymbol{\eta})\, \omega_{\rm lens} \\ \boldsymbol{\phi}_{\rm nz\,WL} &= \boldsymbol{\phi}_{\rm nz}\, \omega_{\rm lens}\,. \end{aligned} \tag{A7}$$

The new estimating equation $\boldsymbol{\psi}_{\rm WL}([\boldsymbol{\phi}_{\rm nz\,WL}, \boldsymbol{\eta}], \mathbf{f})$ is now adjusted for the lensing weights and can be used in conjunction with the empirical likelihood framework to define a likelihood on the mean $E_{\mathbf{f}}\left[w_{\rm i,WL}(\mathbf{f}, \boldsymbol{\eta})\right]$ in equation (A3). We reiterate that $E_{\mathbf{f}}\left[w_{\rm i,WL}(\mathbf{f}, \boldsymbol{\eta})\right]$ denotes here the expectation over the $w_{\rm i,WL}(\mathbf{f}, \boldsymbol{\eta})$ corresponding to bin i over all galaxies in the sample.

The empirical likelihood framework is a non-parametric approach to estimation, which imposes an empirical discrete distribution over the weights $\boldsymbol{w}_{\mathbf{WL}}(\mathbf{f}, \boldsymbol{\eta})$ and then utilizes Lagrange multipliers to constrain this distribution such that the discrete probabilities sum to unity, are positive, and the estimating function relation

$$E_{\mathbf{f}}\left[\boldsymbol{\psi}_{\rm WL}([\boldsymbol{\phi}_{\rm nz\,WL}, \boldsymbol{\eta}], \mathbf{f})\right] = 0 \tag{A8}$$

is fullfilled. One can show in analogy to Owen (2001) that a profile log-likelihood on the mean equation (A3) is obtained by finding the roots to

$$g(\boldsymbol{\lambda}) = \sum_{i=1}^{N_{\rm gal}} \left(\frac{\boldsymbol{w}_{\rm WL}(\mathbf{f}_i, \boldsymbol{\eta}) - \boldsymbol{\phi}_{\rm nz\,WL}}{N_{\rm gal} - \boldsymbol{\lambda}^T \left(\boldsymbol{w}_{\rm WL}(\mathbf{f_i}, \boldsymbol{\eta}) - \boldsymbol{\phi}_{\rm nz\,WL}\right)} \right)\,, \tag{A9}$$

and subsequently evaluating the profile log-likelihood as

$$\ell([\boldsymbol{\eta}, \boldsymbol{\phi}_{\rm nz\,WL}]) = -\sum_{i=1}^{N_{\rm gal}} \log\left(N_{\rm gal} - \boldsymbol{\lambda}^T(\boldsymbol{w}_{\mathbf{WL}}(\mathbf{f}_i, \boldsymbol{\eta}) - \boldsymbol{\phi}_{\rm nz\,WL})\right)\,. \tag{A10}$$

Equation (A9) is monotonic in $\boldsymbol{\lambda}$, which is a Lagrange multiplier of dimension $N_{\rm bins}$. Here, $N_{\rm gal}$ denotes the number of galaxies in the sample. We reach a root for $\boldsymbol{\lambda} = 0$, where $\boldsymbol{\phi}_{\rm nz\,WL} = \frac{1}{N_{\rm gal}} \sum_{i=1}^{N_{\rm gal}} \boldsymbol{w}_{\rm WL}(\mathbf{f}_i, \boldsymbol{\eta})$. This corresponds to the empirical mean of the weights $\boldsymbol{w}_{\rm WL}(\mathbf{f}, \boldsymbol{\eta})$, often referred to as the 'stacked distribution'. This terminology is conventional but misleading because it is often applied inappropriately to summing up likelihood functions of forward models, which is an undefined operation. We refer to Appendix B for a discussion on estimating the $p_{\rm samp}(z)$ in this context.

The central limit theorem holds for the empirical likelihood framework and the coverage error converges as $1/N$, where N denotes the sample size (Owen 2001). Thus, for the large sample sizes considered in this work, we can safely neglect the statistical error in the maximum empirical likelihood estimate, given that other error contributions, such as model misspecification error and cosmic variance, are considerably larger.

APPENDIX B: FORWARD MODELLING METHODOLOGY

B1 Overview

In this appendix, we describe the forward modelling formulation of sample redshift inference in more detail and derive a variational inference scheme to perform efficient $p_{\rm samp}(z)$ inference in this framework. In Section 4.1, we discussed a simplified model, focussing on the redshift z as the quantity of interest, as (e.g. Leistedt et al. 2016; Malz & Hogg 2020; Rau et al. 2022)

$$p(\hat{\mathbf{F}}|\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\Omega}) = \prod_{i=1}^{N_{\rm gal}} \int {\rm d}z_i \; \omega_i \; p(\mathbf{f}_i|z_i, \boldsymbol{\Omega}) p(z_i|\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\Omega}) \,. \tag{B1}$$

We reiterate that $\hat{\mathbf{F}}$ denotes the set of fluxes of all $N_{\rm gal}$ galaxies in the sample, $\mathbf{f}_i\,(z_i)$ denotes the flux (redshift) of the individual galaxy with index i, and $\boldsymbol{\Omega}$ denotes a set of auxiliary parameters that describe other galaxy properties such as galaxy type or stellar mass. The weights ω_i denote the lensing weights for each galaxy in the sample.

B2 Assumptions

The simplified equation (B1) assumes that the flux and redshift of each galaxy are drawn independently of any other. In a more general setting, we could formulate a joint likelihood. The forward modelling approach does not assume the availability of calibration data and is therefore more general than the conditional density estimation methodology. In contrast to conditional density estimation, equation (B1) implies a hierarchical inference of the $p_{\rm samp}(z)$. The same applies to other population distributions for quantities of interest. For noisy measurements of photometry, this inverse problem can be poorly conditioned. Practical applications must impose explicit or implicit assumptions to control the posterior variance, either by setting priors on quantities of interest or restricting the complexity of relevant models. Model misspecification error is a significant complication in this context. Given the complex modelling of SEDs, selection functions, and photometric error, any practical application must verify their modelling assumptions on calibration data.

B3 Methodology

We discretize the $\vec{p}_{\rm indiv}(z)$ on the same grid that defines the $p_{\rm samp}(z)$ histogram defined in equation (1). We define a matrix defined as the set:

$$\mathbf{pz} := \{pz_{\rm ij}(\boldsymbol{\Omega}) \,|\, 0 < i \leq N_{\rm gal}, 0 < j \leq N_{\rm bins}\} \,, \tag{B2}$$

where the entries are given as the integrals of the likelihood of galaxy i over the j redshift histogram bin weighted by the lensing weights ω_i

$$pz_{\rm ij}(\boldsymbol{\Omega}) := \omega_i \int p(\mathbf{f}_i|z_i, \boldsymbol{\Omega}) \mathbb{1}_j(z) {\rm d}z_i \,. \tag{B3}$$

Using the definition equation (1) we can write the log-likelihood as

$$\log\left(p(\hat{\mathbf{F}}|\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\Omega})\right) = \sum_{i=1}^{N_{\rm gal}} \log\left(\sum_{j}^{N_{\rm bins}} \phi_{\rm nz,j} pz_{\rm ij}(\boldsymbol{\Omega})\right) . \tag{B4}$$

The logarithm in equation (B4) and the fact that $\boldsymbol{\phi}_{\rm nz}$ is normalized (it can be transformed to lie on the simplex) makes the evaluation and optimization of equation (B4) non-trivial.[20]

We can circumvent both issues by introducing the binary variables ρ_{ij} that associate bin j with galaxy i. The complete data likelihood then reads

$$p(\hat{\mathbf{F}}, \boldsymbol{\rho}|\boldsymbol{\phi}_{\rm nz}, \boldsymbol{\Omega}) \propto \prod_{i=1}^{N_{\rm gal}} \prod_{j=1}^{N_{\rm bins}} \left(\phi_{\rm nz,j}\, pz_{\rm ij}(\boldsymbol{\Omega})\right)^{\rho_{\rm ij}} , \tag{B5}$$

which we identify as a multinominal likelihood. Imposing a Dirichlet prior over the parameters $\boldsymbol{\phi}_{\rm nz}$ then yields the joint distribution $p(\hat{\mathbf{F}}, \boldsymbol{\rho}, \boldsymbol{\phi}_{\rm nz}|\boldsymbol{\Omega})$.

Variational inference maximizes the Evidence Lower Bound (ELBO), which is equivalent to minimizing the Kullback–Leibler divergence between the true, unknown, posterior, and an 'ansatz', the variational distribution

$${\rm ELBO} = E_{q(\boldsymbol{\rho}, \boldsymbol{\phi}_{\rm nz})}\left[\log\left(p(\hat{\mathbf{F}}, \boldsymbol{\rho}, \boldsymbol{\phi}_{\rm nz}|\boldsymbol{\Omega})\right) - \log q(\boldsymbol{\rho}, \boldsymbol{\phi}_{\rm nz})\right] , \tag{B6}$$

where $q(\boldsymbol{\rho}, \boldsymbol{\phi}_{\rm nz})$ denotes the variational distribution to be optimized. Here, this involves imposing an analytic form for the variational distribution and then maximizing the ELBO with respect to its parameters.

We make a mean-field ansatz for the variational distribution

$$q(\boldsymbol{\rho}, \boldsymbol{\phi}_{\rm nz}) \approx q(\boldsymbol{\rho}) q(\boldsymbol{\phi}_{\rm nz}) \,, \tag{B7}$$

which assumes independence between $\boldsymbol{\rho}$ and $\boldsymbol{\phi}_{\rm nz}$.

Under the mean-field approximation, variational inference reduces to a simple scheme of updating each component iteratively by mean-field coordinate ascent. Setting the Lagrange function constructed using the variational derivative of the ELBO to zero, we can derive the following coordinate ascent iteration steps:

$$\begin{aligned} q(\boldsymbol{\rho}) &\propto \exp\left(E_{q(\boldsymbol{\phi}_{\rm nz})}\left[\log p(\boldsymbol{\rho}|\boldsymbol{\phi}_{\rm nz}, \mathbf{pz})\right]\right) \\ &\propto \prod_{i=1}^{N_{\rm gal}} \prod_{j=1}^{N_{\rm bins}} \left(\exp\left(\psi(\alpha_j) - \psi\left(\sum_{a=1}^{N_{\rm bins}} \alpha_a\right) + \log\left(pz_{ij}\right)\right)\right)^{\rho_{ij}} , \end{aligned} \tag{B8}$$

and

$$\begin{aligned} q(\boldsymbol{\phi}_{\rm nz}) &\propto \exp\left(E_{q(\boldsymbol{\rho})}\left[\log p(\boldsymbol{\phi}_{\rm nz}|\boldsymbol{\rho}, \mathbf{pz})\right]\right) \\ &= {\rm Dir}(\boldsymbol{\alpha_0} + \sum_{i=1}^{N_{\rm gal}} \boldsymbol{\gamma_i}), \end{aligned} \tag{B9}$$

where we have omitted the conditioning of the variational distributions on the parameter $\boldsymbol{\alpha}$ for notational convenience. We note that $\boldsymbol{\alpha}$ is iteratively updated in the argument of the Dirichlet defined in equation (B9). The sum in equation (B9) goes over the $N_{\rm gal}$-dimension of the matrix, whose elements are defined as

$$\gamma_{ij} = \frac{\exp\left(\psi(\alpha_j) - \psi(\sum_{a=1}^{N_{\rm bins}} \alpha_a) + \log\left(pz_{\rm ij}\right)\right)}{\sum_{j=1}^{N_{\rm bins}} \exp\left(\psi(\alpha_j) - \psi(\sum_{a=1}^{N_{\rm bins}} \alpha_a) + \log\left(pz_{\rm ij}\right)\right)} . \tag{B10}$$

Here, ψ denotes the digamma function; the Dirichlet distribution is abbreviated as 'Dir'. The variational distributions defined in equation (B8) and equation (B9) are iteratively updated until convergence.

[20] A possible way to perform the optimization in a brute-force approach is by projected gradient descent. However, we derive a simpler scheme in the following text.

While this iterative scheme can be expected to computationally outperform MCMC approaches, a mean-field ansatz often leads to the estimation of too narrow credibility intervals.

In our numerical experience, the undercoverage[21] under reasonable regularization (e.g. by selecting broader histogram bins) is approximately 20 per cent, which is subdominant compared with other sources of error induced by spatial-, colour- and redshift-dependent selection functions or model misspecification. We therefore used the variational inference scheme in this work during the initial stages of the project, where we evaluated the accuracy of the Mizuki individual galaxy photometric redshifts. However, we note that the validity of the variational inference approximation will depend on the resolution (e.g. given by the histogram bins size) and can be expected to deteriorate for poorly conditioned scenarios with high variance. In these cases, we can expect credibility intervals to exhibit undercoverage. In contrast, maximum a posteriori predictions can be expected to be still entirely accurate.

APPENDIX C: CHARACTERIZING COSMIC VARIANCE USING LOGISTIC GAUSSIAN PROCESSES

In this appendix, we discuss how logistic Gaussian processes provide a flexible model to include cosmic variance induced sample noise into $p_{\rm samp}(z)$ inference.

We first consider the redshift-dependent lognormal doubly stochastic point process specified as

$$\begin{aligned} \rho_i &\sim \mathrm{LogNorm}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ N_i &\sim \mathrm{Poisson}(\rho_i)\,, \end{aligned} \tag{C1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance parameters of the lognormal distribution, the ρ_i are the mean parameters of the Poisson distribution that describes the galaxy number counts in redshift dimension, and N_i denotes the number of galaxies in each redshift bin. The $p_{\rm samp}(z)$, which enters the modelling of two-point statistics, is normalized to integrate to unity. We therefore need to sample over normalized histogram counts of a multinomial instead of parameters of a Poisson distribution.

The lognormal 'Cox process' defined in equation (C1) can be equivalently defined as

$$\begin{aligned} \rho_i &\sim \mathrm{LogNorm}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \mathbf{N} &\sim \mathrm{MultNominal}(\boldsymbol{\phi}_{\rm nz}, \overline{N}) \\ \overline{N} &\sim \mathrm{Poisson}(\overline{\rho}) \\ \phi_{{\rm nz},i} &= \frac{\rho_i}{\sum_{i=0}^{N_{\rm bins}} \rho_i} \\ \overline{\rho} &= \sum_{i=0}^{N_{\rm bins}} \rho_i\,, \end{aligned} \tag{C2}$$

where we have decomposed the Poisson distribution into the product of a multinominal distribution that depends on the normalized $\boldsymbol{\rho}$-parameters and a Poisson distribution that depends on their sum $\overline{\rho}$. The random variable $\overline{N}$ denotes the total number of galaxies across all bins.

Here, $N_{\rm bins}$ is the number of redshift bins introduced in equation (1). Since the modelling of the angular correlation function depends on the normalized $p_{\rm samp}(z)$, we will concentrate on the distribution of $\boldsymbol{\phi}_{\rm nz}$, where equation (C2) defines the logistic Gaussian process specification of our model. We make the simplifying assumption of ignoring the cross-correlations between neighbouring redshift bins, which has been shown to be a reasonable approximation in Sánchez et al. (2020). To include an error contribution to the lognormal model covariance that matches the variation in the COSMOS field, we are interested to predict the coefficient of variation, i.e. the ratio between the standard deviation and mean, for the HSC phot data in the COSMOS field as a function of redshift due to cosmic variance and use it to derive a cosmic variance error budget on the $p_{\rm samp}(z)$ model for each tomographic bin. To this end, we first formulate a model for the variance of galaxy counts as a function of redshift that can be fitted to the results of Sánchez et al. (2020).

Consider two sets of galaxies within a spatial area and redshift bin, which we denote as B_1 and B_2. We can express the covariance of the number of galaxies within the sets $N(B_{1/2})$ as (e.g. Stoyan & Stoyan 1994)

$$\begin{aligned} &\mathrm{cov}(N(B_1), N(B_2)) = E\left[N(B_1)N(B_2)\right] - E\left[N(B_1)\right] E\left[N(B_2)\right] \\ &= E\left[\sum_{\mathbf{x_1}\in N(B_1)} \sum_{\mathbf{x_2}\in N(B_2)} \mathbb{1}_{B_1}(\mathbf{x_1})\mathbb{1}_{B_2}(\mathbf{x_2})\right] - \rho_{B_1} V(B_1)\rho_{B_2} V(B_2)\,, \end{aligned} \tag{C3}$$

where $V(B_{1/2})$ and $\rho_{B_{1/2}}$ denotes the volume and expected number density of B_1 and B_2. The volume is defined with respect to spatial area and redshift bin and the expected number density ρ denotes the expected number of galaxies observed in B per unit volume. $\mathbb{1}_B(\mathbf{x})$ denotes the indicator function which is unity if a galaxy can be found at position $\mathbf{x}$ and zero otherwise. The first term corresponds to the second-moment measure, i.e. the expected number of galaxy pairs including 'pairs' of the same galaxy. This can be expressed as a function of the two-point correlation function, the number densities, and the effect of the survey mask. The variance contribution we obtain within a set B under the assumption of homogeneity and isotropy can be defined as

$$Var[N(B)] = \rho_B V(B) + \rho_B^2 \iint_B \xi(||\mathbf{x_1} - \mathbf{x_2}||) d\mathbf{x_1} d\mathbf{x_2}\,. \tag{C4}$$

The first term in equation (C4) is the 'shot noise' contribution. The second term in equation (C4) depends on the 'clustering' of the galaxy field, parametrized by the pair-correlation function $\xi(x_1, x_2)$ and the survey geometry that enters the double integral over B.

We develop a simple model for the COSMOS2015 data based on equation (C4) by parametrizing ρ_B proportional to a lognormal distribution and the integral of the correlation function proportional to a power law. Our model has five parameters; an amplitude and scale parameter for the power-law model and two parameters that describe the line-of-sight number density of the COSMOS2015 number counts with a normalization amplitude. We then fit this model to the redshift-dependent $Var[N]/N$ values reported in Sánchez et al. (2020), shown in the left panel of Fig. C1. The black dashed line shows the values reported in Sánchez et al. (2020), and the red line shows the best-fitting solution to our model. We see that at low redshift the linear dependence on the lognormal-shaped line-of-sight number density of the COSMOS2015 number counts flattens the power-law shape. In the right panel of Fig. C1, we plot the coefficient of variation (red) and the coefficient of variation[22] from only the shot noise contribution, i.e. the first term of equation (C4). In agreement

[21] Undercoverage refers here to underestimating the width of the credibility intervals.

[22] The coefficient of variation is the ratio of the standard deviation to the mean.

with Sánchez et al. (2020) we see that the shot noise contribution is subdominant for the COSMOS2015 data set. This difference will be even larger for our data due to the larger amount of galaxies in HSC phot.

The cosmic variance contribution to the coefficient of variation is strictly bounded from above by the total coefficient of variation by

$$\sqrt{\iint_B \xi(||\mathbf{x_1}-\mathbf{x_2}||)d\mathbf{x_1}d\mathbf{x_2}} \leq \sigma[N(B)]/E[N(B)]\,. \tag{C5}$$

We choose to use the 'full' coefficient of variation from COSMOS2015 (CV), in our model, even though the shot noise contribution would already be included in the empirical likelihood framework (or the deconvolution approach in the Mizuki case), which will lead to an overestimation of our error budget following equation (C5).

In order to derive the cosmic variance error contribution on the redshift distribution, we scale the CV by the number counts in redshift bins as predicted by the empirical likelihood framework (or alternatively by our deconvolution algorithm) using

$$E[N_i] = N_{\rm tot}\,\pi_i^{\rm ML}\,, \tag{C6}$$

where $\pi_i^{\rm ML}$ defines the maximum empirical likelihood estimate in redshift bin i as discussed in Section 5.4, and $(N_i/N_{\rm tot})$ denotes the (redshift bin i/total number of galaxies) in the tomographic bin.

Using the method of moments we can now estimate the parameters $\boldsymbol{\mu}$ and Σ defined in equation (C2) as

$$\begin{aligned}\mu_i &= \log\left(\frac{E[N_i]}{\sqrt{\mathrm{CV}_i^2+1}}\right)\\ \sigma_i^2 &= \log\left(\mathrm{CV_i}^2+1\right),\end{aligned} \tag{C7}$$

where the coefficient of variation is given as

$$\mathrm{CV_i} = \sigma[N_i]/E[N_i]\,. \tag{C8}$$

This allows us to specify the logistic Gaussian process prior in equation (C2) defined in $N_{\rm bins}$ redshift bins for each of the four tomographic bins in our sample.

Given these definitions we can simplify the specification of the logistic Gaussian process on the parameters $\boldsymbol{\phi}_{\rm nz}$ in equation (1) to

$$\begin{aligned}\boldsymbol{s} &\sim N(\boldsymbol{s}|\mu, \Sigma_{\rm CV})\\ \boldsymbol{\phi}_{\rm nz} &:= \left\{\frac{\exp(s_i)}{\sum_j \exp(s_j)}\,\middle|\, 0<i<N_{\rm bins}\right\},\end{aligned} \tag{C9}$$

where $\boldsymbol{\mu}$ and the diagonal matrix $\Sigma_{\rm CV}$ are defined in equation (C7). The sampling of the $\boldsymbol{\phi}_{\rm nz}$ parameters is expressed in terms of the variable $\mathbf{s}$ that follows a multivariate normal distribution. This corresponds to the definition in equation (6).

APPENDIX D: MARGINALIZING OVER THE CAMIRA LRG PHOTOMETRIC REDSHIFT ERROR

In the following text, we describe the definition of the marginal likelihood that accounts for the photometric redshift error of the CAMIRA LRG (LRG) sample introduced in Section 3.2. In this approach we treat the redshifts of each LRG as a latent variable. Since we do not have access to the likelihood of the photometric redshift method implemented in the CAMIRA method, we utilize the calibration data set described in Section 3.3 to estimate a conditional distribution between the flux of the LRGs $\mathbf{f}_{\rm LRG}$ and their redshift $z_{\rm LRG}$. This is done by matching the LRG catalogue and the specXphot calibration catalogue and constructing a kernel based conditional density estimate. We can then marginalize the likelihood of spatial cross-correlations between the LRG and HSC photometric sample (phot) as

$$\begin{aligned}&p(\hat{\mathbf{w}}_{\rm LRG-PhotZ}|\boldsymbol{\phi}_{\rm nz\,PhotZ}, \mathbf{b}_{\rm PhotZ}, \mathbf{b}_{\rm LRG})\\ &= \iint p(\hat{\mathbf{w}}_{\rm LRG-PhotZ}|\boldsymbol{\phi}_{\rm nz\,PhotZ}, \mathbf{b}_{\rm PhotZ}, \mathbf{b}_{\rm LRG}, \mathbf{z}_{\rm LRG}, \mathbf{f}_{\rm LRG})\\ &\times p(\mathbf{z}_{\rm LRG}|\mathbf{f}_{\rm LRG})p(\mathbf{f}_{\rm LRG})\,d\mathbf{f}_{\rm LRG}\,d\mathbf{z}_{\rm LRG}\,,\end{aligned} \tag{D1}$$

where $\hat{\mathbf{w}}_{\rm LRG-PhotZ}$ denotes the spatial cross-correlation measurements between the LRG and HSC phot catalogues, $\boldsymbol{\phi}_{\rm nz\,PhotZ}$ denotes the $p_{\rm samp}(z)$ parameters of the HSC phot sample, and $(\mathbf{b_{PhotZ}}/\mathbf{b_{LRG}})$ is the galaxy-dark matter bias of the (HSC phot/CAMIRA LRG) sample. The left-hand side defines the marginal likelihood introduced in equation (8). The term $p(\mathbf{z_{LRG}}|\mathbf{f}_{\rm LRG})$ is the aforementioned conditional distribution of the LRGs' redshift given their flux. We also include the lensing weights for the HSC phot sample by weighting the pair counts used to construct the measurement $\hat{\mathbf{w}}_{\rm LRG-PhotZ}$ according to the prescription implemented in 'The-Wizz' (Morrison et al. 2017).

Since the cross-correlation measurements do not vary much between realizations of LRG redshifts drawn from $p(\mathbf{z_{LRG}}|\mathbf{f}_{\rm LRG})$, we can evaluate this double integral using a Monte Carlo estimate:

$$\begin{aligned}&\hat{p}(\hat{\mathbf{w}}_{\rm LRG-PhotZ}|\boldsymbol{\phi}_{\rm nz\,PhotZ}, \mathbf{b}_{\rm PhotZ}, \mathbf{b}_{\rm LRG})\\ &= \frac{1}{M}\sum_{(\mathbf{f}_{\rm LRG}, \mathbf{z}_{\rm LRG})} \left(p(\hat{\mathbf{w}}_{\rm LRG-PhotZ}|\boldsymbol{\phi}_{\rm nz\,PhotZ}, \mathbf{b}_{\rm PhotZ}, \mathbf{b}_{\rm LRG}, \mathbf{z}_{\rm LRG}, \mathbf{f}_{\rm LRG})\right)\,,\end{aligned} \tag{D2}$$

where we sample M sets $(\mathbf{f_{LRG}}, \mathbf{z_{LRG}})$ from the estimated joint distribution $\hat{p}(\mathbf{f}_{\rm LRG}, \mathbf{z}_{\rm LRG})$ by sampling sequentially as

$$\begin{aligned}\mathbf{f}_{\rm LRG} &\sim \hat{p}(\mathbf{f}_{\rm LRG})\\ \mathbf{z}_{\rm LRG} &\sim \hat{p}(\mathbf{z}_{\rm LRG}|\mathbf{f}_{\rm LRG})\,.\end{aligned} \tag{D3}$$

In this sampling scheme one has to recalculate the lensing-weighted pair-counts for each replication. This has the advantage that the scales and redshift bins can be consistently selected, but the disadvantage of high computational cost. However, we iterate and verify that the variance in the integrand is moderate due to the small LRG photometric redshift error. Accordingly, we can use a small number of realizations ($M = 10$ in our case), which makes this a practical approach. We finally note that we speed up the construction of the conditional density estimate $\hat{p}(\mathbf{z}_{\rm LRG}|\mathbf{f}_{\rm LRG})$ by training directly on the residuals between the specXphot 'true' redshifts in the training set $\mathbf{z_{LRG}}$ and the estimated mean photometric redshift estimates $\mathbf{z}_{\rm phot,\,LRG}$, i.e. we construct $\hat{p}(\mathbf{z}_{\rm LRG}|\mathbf{z}_{\rm phot,LRG})$.

While this potentially increases the variation in the resampled CAMIRA LRG redshifts, since we do not use the full information in the photometry as predictors, it allows us to train our error model efficiently on subsamples of LRG galaxies with very small variations between the conditional density function estimates due to the higher density of LRG training galaxies in the one-dimensional covariate $\mathbf{z}_{\rm phot,\,LRG}$.

APPENDIX E: DISCUSSION OF PRIOR CHOICE

In this appendix, we discuss methodological differences between the logistic Gaussian Process as a prior over the $p_{\rm samp}(z)$ and the established alternative choice of the Dirichlet.

We have introduced the logistic Gaussian process as a prior distribution over $p_{\rm samp}(z)$ in Rau et al. (2020), where we discuss several advantages in terms of characterizing the covariance between

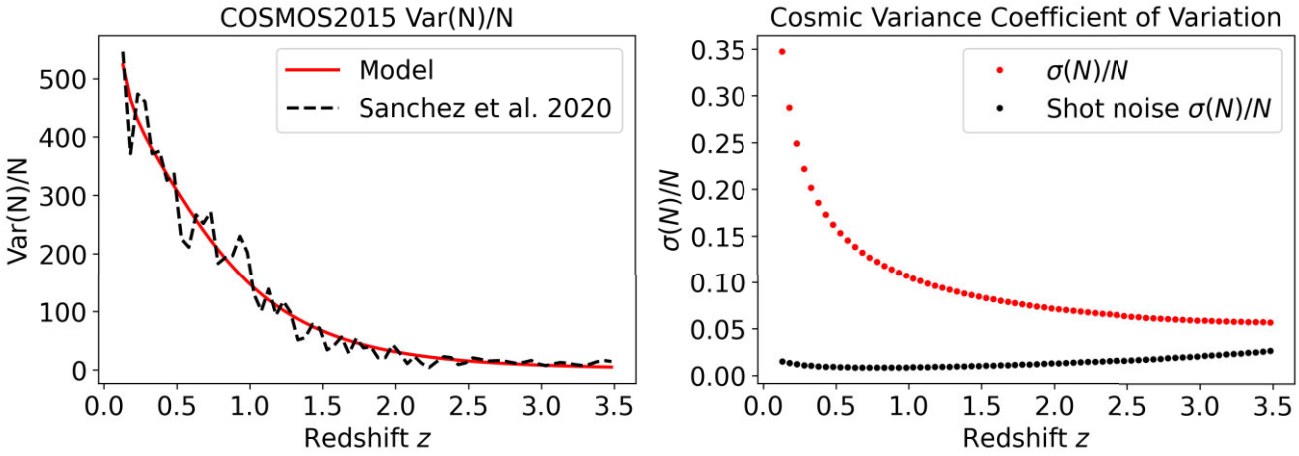

Figure C1. Left: *Var*(*N*)/*N* model for the COSMOS2015 data as a function of redshift used in this work (red) compared with the (black, dashed) predictions in Sánchez et al. (2020). Right: Coefficient of variation as a function of redshift predicted by our model (red) compared with the contribution from shot noise alone (black).

neighbouring redshift bins. Furthermore, as explained in Appendix C we can relate our choice of logistic Gaussian process prior to a lognormal model for the one-point density along the line of sight.

The Dirichlet distribution is an alternative prior that can be imposed over coefficients of finite basis function models like e.g. the histogram. It is a conjugate prior to the multinomial likelihood which is a significant advantage in designing sampling and inference schemes as, for example, demonstrated in the derivation of the variational inference scheme in Appendix B. In this context it is often applied as an uninformative prior over the histogram heights.

The Dirichlet distribution is related to a gamma distribution Gamma($\boldsymbol{\alpha}$, 1) in a similar way as the logistic Gaussian Process to the lognormal model.

$$\begin{aligned} \rho_i &\sim \text{Gamma}(\boldsymbol{\alpha}, 1) \\ \mathbf{N} &\sim \text{MultNominal}(\boldsymbol{\phi}_{\text{nz}}, \overline{N}) \\ \phi_{\text{nz,i}} &= \frac{\rho_i}{\sum_{i=0}^{N_{\text{bins}}} \rho_i}, \end{aligned} \tag{E1}$$

where the vector $\mathbf{N}$ denotes the galaxy counts drawn from the multinomial and $\overline{N}$ denotes the total number of galaxies. The vector $\boldsymbol{\phi}_{\text{nz}}$ would then be distributed according to a Dirichlet distribution with coefficients $\boldsymbol{\alpha}$.

Concentrating on the distribution of $\boldsymbol{\rho}$, which describes the expected number density of the point process along the line of sight, the logistic Gaussian process as the prior over the $p_{\text{samp}}(z)$ implies a lognormal model, dependent on both a mean vector and covariance, whereas the choice of a Dirichlet distribution implies a one parameter Gamma distribution, dependent on the vector $\boldsymbol{\alpha}$. The limitation of the one-parameter Gamma distribution is that both the coefficient of variation and the average number density depend on the same parameter vector $\boldsymbol{\alpha}$. This means that we cannot parametrize a redshift-dependency in the coefficient of variation in the Dirichlet model while leaving the mean histogram heights constant. We can however change the average coefficient of variation while leaving the mean constant as demonstrated in the following. Following Minka (2000) we can reparametrize the Dirichlet as

$$\begin{aligned} s &= \sum_k \alpha_k \\ \boldsymbol{m} &= \frac{\boldsymbol{\alpha}}{s}, \end{aligned} \tag{E2}$$

where ($\boldsymbol{m}$/s) relates to the (mean/precision) of the Dirichlet distribution over the histogram heights. When the mean $\mathbf{m}$ is kept constant, one can modify the standard deviation of the Dirichlet distribution by scaling the precision s. Sánchez et al. (2020) mention this aspect in their work in a slightly different context. We further note that typically the lognormal distribution can be adjusted to be close to the Gamma distribution.

In summary, we use the logistic Gaussian Process model in this work as it allows a more flexible parametrization of uncertainty compared with the Dirichlet model.

This paper has been typeset from a TEX/LATEX file prepared by the author.

Published by Oxford University
Press on behalf of Royal Astronomical Society 2023. This work is written by (a) US Government employee(s) and is in the public domain in the US.