Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **507,** 3187–3206 (2021)
Advance Access publication 2021 July 16

https://doi.org/10.1093/mnras/stab1986

Correcting correlation functions for redshift-dependent interloper contamination

Daniel J. Farrow [1,2]★ Ariel G. Sánchez [1,2] Robin Ciardullo,[3,4] Erin Mentuch Cooper [5] Dustin Davis [5] Maximilian Fabricius [1,2] Eric Gawiser,[6] Henry S. Grasshorn Gebhardt,[7,8] Karl Gebhardt,[5] Gary J. Hill,[5,9] Donghui Jeong,[3,4] Eiichiro Komatsu,[10,11] Martin Landriau,[12] Chenxu Liu,[5] Shun Saito [11,13] Jan Snigula[1,2] and Isak G. B. Wold[14]

[1]*Max-Planck-Institut für extraterrestrische Physik, Giessenbachstrasse 1, D-85748 Garching, Germany*
[2]*Fakultät für Physik, Universitäts-Sternwarte, Ludwig-Maximilians-Universität München, Scheinerstr. 1, D-81679 München, Germany*
[3]*Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, PA 169802, USA*
[4]*Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, PA 169802, USA*
[5]*Department of Astronomy, University of Texas at Austin, 2515 Speedway, Stop C1400, Austin, TX 78712, USA*
[6]*Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA*
[7]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA*
[8]*California Institute of Technology, Pasadena, CA 91125, USA*
[9]*McDonald Observatory, University of Texas at Austin, 2515 Speedway, Stop C1402, Austin, TX 78712, USA*
[10]*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild Str. 1, D-85741 Garching, Germany*
[11]*Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU, WPI), University of Tokyo, Chiba 277-8582, Japan*
[12]*Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[13]*Institute for Multi-messenger Astrophysics and Cosmology, Department of Physics, Missouri University of Science and Technology, 1315 N Pine St, Rolla, MO 65409, USA*
[14]*Astrophysics Science Division, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA*

Accepted 2021 July 2. Received 2021 June 29; in original form 2021 April 9

ABSTRACT

The construction of catalogues of a particular type of galaxy can be complicated by interlopers contaminating the sample. In spectroscopic galaxy surveys this can be due to the misclassification of an emission line; for example in the Hobby–Eberly Telescope Dark Energy Experiment (HETDEX) low-redshift [O II] emitters may make up a few per cent of the observed Ly α emitter (LAE) sample. The presence of contaminants affects the measured correlation functions and power spectra. Previous attempts to deal with this using the cross-correlation function have assumed sources at a fixed redshift, or not modelled evolution within the adopted redshift bins. However, in spectroscopic surveys like HETDEX, where the contamination fraction is likely to be redshift dependent, the observed clustering of misclassified sources will appear to evolve strongly due to projection effects, even if their true clustering does not. We present a practical method for accounting for the presence of contaminants with redshift-dependent contamination fractions and projected clustering. We show using mock catalogues that our method, unlike existing approaches, yields unbiased clustering measurements from the upcoming HETDEX survey in scenarios with redshift-dependent contamination fractions within the redshift bins used. We show our method returns autocorrelation functions with systematic biases much smaller than the statistical noise for samples with at least as high as 7 per cent contamination. We also present and test a method for fitting for the redshift-dependent interloper fraction using the LAE–[O II] galaxy cross-correlation function, which gives less biased results than assuming a single interloper fraction for the whole sample.

Key words: methods: data analysis – cosmology: observations – large-scale structure of the Universe.

1 INTRODUCTION

The measurement of a redshift from a galaxy spectrum is one of the most fundamental parts of a spectroscopic survey. This is usually achieved by relying on features in the spectra such as emission and absorption lines and the shape of the continuum. However, when only one emission line is detected, it becomes impossible to unambiguously identify the rest-frame emission line and return an accurate classification and redshift. This results in catalogues of galaxies that contain interlopers, i.e. misclassified sources at the wrong redshift. Interloper contamination is expected to be important in several major upcoming galaxy surveys (see e.g. Pullen et al. 2016). The focus of this paper is the ongoing Hobby–Eberly Telescope Dark Energy Experiment (HETDEX; Hill et al. 2008, in preparation; Gebhardt et al. in preparation), where, due to the spectrographs not resolving the [O II] doublet, low-redshift [O II]

★ E-mail: dfarrow@mpe.mpg.de

emitters with rest-frame wavelength 3727 Å can be mistaken for high redshift Ly α emitters (LAEs) with rest-frame wavelength 1216 Å.

The impact of interlopers on the correlation function and power spectrum of a galaxy sample has been studied in the literature (e.g. Pullen et al. 2016; Leung et al. 2017; Addison et al. 2019; Grasshorn Gebhardt et al. 2019; Massara et al. 2020). It has been seen that the presence of interlopers in a sample changes the galaxies' correlation function and power spectrum. It is also understood that if the interlopers are unclustered then the main effect just decreases the overall clustering amplitude by adding in uncorrelated sources (see Appendix B.4 of Grasshorn Gebhardt et al. 2019). However, if the interlopers are clustered, then a signal from their correlation function is added into the sample. It has also been shown that these spurious clustering signals can cause biases in the inferred cosmological parameters (e.g. Pullen et al. 2016; Addison et al. 2019; Grasshorn Gebhardt et al. 2019).

In both Addison et al. (2019) and Grasshorn Gebhardt et al. (2019), methods are presented that include the effects of interlopers in the modelling of the galaxy power spectrum. These authors note that a cross-correlation signal between two intrinsically uncorrelated samples of galaxies can be created entirely due to interloper contamination. They advocate using this observed cross-correlation signal to put constraints on the contamination fraction, in order to yield better measurements of cosmological parameters. An alternative approach to forward modelling techniques is to decontaminate the measurements by applying a transformation that changes the observed auto and cross-correlation functions into the true underlying functions. A matrix to carry out this transformation and its inverse is given in Awan & Gawiser (2020). Their work deals with angular clustering measurements in redshift bins.

A related issue to interlopers in spectroscopic galaxy surveys is their impact in line intensity mapping experiments (e.g. Visbal & Loeb 2010; Gong et al. 2014; Cheng et al. 2016; Lidz & Taylor 2016; Cheng, Chang & Bock 2020; Gong, Chen & Cooray 2020). These studies differ from emission line surveys in that they target the light from unresolved populations of galaxies. However, it has also been noted that interlopers in intensity mapping experiments add an anisotropic signal to the power spectrum of the target population (e.g. Visbal & Loeb 2010; Gong et al. 2014; Lidz & Taylor 2016). In Gong et al. (2020), a method is presented that jointly fits the cosmology and properties of interloper lines in line intensity mapping experiments.

One scenario that has not been addressed by efforts to model the correlation function or power spectrum from spectroscopic emission line surveys is when the contamination fractions and the clustering of the contaminants show rapid evolution within the redshift bins used to define samples. Existing methods may work to an acceptable level with correlation functions that have a reasonable amount of evolution within the redshift bins considered, but in HETDEX, the observed [O II] clustering signal will evolve rapidly with redshift, due to projection effects (see e.g. fig. 2 of Grasshorn Gebhardt et al. 2019). The [O II] contamination fraction will also be redshift dependent, due to the intrinsic redshift distribution of the emission lines and due to the wavelength dependence of the noise. Although Cheng et al. (2020) recently published a method of generating a 3D lightcone of the interlopers in an intensity mapping survey, their method relies on the interlopers having multiple emission lines. That will not usually be the case for HETDEX, as beyond $z \sim 0.13$, the bulk of the [O II] galaxy population will only have a single detectable emission line. Cheng et al. (2020) also focus on producing a 3D map of the interloper density, not unbiased correlation function measurements from the target population.

In this paper, we present a method to account for the redshift dependence of the contamination fractions in emission-line surveys by combining the decontamination methodology in the literature with lightcone effects presented in Yamamoto & Suto (1999) and Suto, Magira & Yamamoto (2000). References to 'lightcone effects' in this paper specifically refer to effects from the redshift dependent contamination and observed clustering. We test our method on simulations of the HETDEX survey, and demonstrate that our method to deal with the lightcone effects is an improvement over assuming fixed contamination fractions and clustering across a whole redshift bin. We also show that our new method is useful when using the cross-correlation function to gain unbiased constraints on the contamination fractions. We focus on HETDEX here, but the work we present gives insights into all surveys with contamination rates that depend on redshift.

The outline of this paper is as follows: in Section 2, we introduce the HETDEX survey and our simulations of it; this section also includes a method of assigning source classification probabilities. In Section 3, we present the methods used to measure and model the projected clustering. Then in Section 4, we present the methodology of our decontamination. We show the results of our model in Section 5, and in Section 6, we use our new methodology to fit for the redshift-dependent contamination. We give our conclusions in Section 7.

2 SIMULATIONS OF HETDEX

In this section, we explain how we generate mock catalogues. We note that our work follows that of Chiang et al. (2013), who use an older version of the lognormal simulation code used here (Agrawal et al. 2017), and an older HETDEX design, to produce simulations of the HETDEX survey. We improve on that paper, first by adding [O II] galaxies and source classifications following Leung et al. (2017), and then by adding in more realistic redshift dependent variations into the sensitivity and noise estimates.

We will begin by introducing HETDEX (Section 2.1), then the following sections introduce the model of large-scale structure (Section 2.2) and the approach we use to generate a density field with a given power spectrum (Section 2.3). We also explain how we assign galaxy properties (Sections 2.5 and 2.9), model observational effects (Sections 2.4, 2.6, and 2.7) and assign the LAE probabilities (Section 2.10) to generate samples of LAEs and [O II] emitters.

2.1 The HETDEX survey

HETDEX is a program on the Hobby–Eberly Telescope at the McDonald Observatory, Texas (Hill et al. 2008, in preparation; Gebhardt et al. in preparation) to use LAEs to map out the large-scale structure of the $1.9 < z < 3.5$ Universe. The survey measures spectra from the sky using an array of up to 78 integral field units (IFUs; Hill et al. 2018, in preparation), galaxies are not pre-selected but instead observations are taken blindly. Each IFU has a square footprint roughly 50 arcsec on a side, and neighbouring IFUs are separated by 100 arcsec. When observing, the gaps between the fibres are filled in by taking three dithered exposures. The dithering does not fill in the gaps between the IFUs, however, meaning areas of sky are sparsely sampled. It has been shown that such a sampling can be treated as surveying the whole area with a lower number of tracers (Chiang et al. 2013). We refer to the set of three dithers at one pointing as an 'exposure set', and use the term 'exposure set position' to refer to the Right Ascension and Declination of the pointing.

The survey sparsely samples two main fields: a roughly 390 deg^2 field in the Northern hemisphere (the 'Spring' field) and an $\sim$150 deg^2 equatorial region (the 'Fall' field). Defining the area that is sparsely sampled is difficult due to the jagged edges of the HETDEX footprint, which are caused by the approximately octagonal boundary of IFUs in the focal plane. In the real survey, additional effects we do not model here, such as bright stars in the Milky Way and large foreground galaxies create holes in the survey, further complicating the issue. Thus, the precise values for the survey areas depend on how survey edges are defined and the regions that are compromised by foreground sources.

The survey goal is to measure the clustering (e.g. correlation function or power spectrum) of the LAEs and use it to probe cosmology. The modest resolution of the spectrographs (mean resolving power $R = \lambda/\delta\lambda \sim 800$) means the [O II] doublet cannot be resolved, resulting in some [O II] emitters being classified as LAEs (see Leung et al. 2017).

2.2 Model of cosmology and large-scale structure

The simulations and our whole paper use the marginalized mean, flat ΛCDM cosmology from the Planck Collaboration VI (2020), but for simplicity we assume massless neutrinos (see table 1 for the exact parameter values). The model of the power spectrum and bias used to generate the simulations is the same as that used for the analysis of the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013) by Sánchez et al. (2017), and a full description of the model can be found there. Briefly, a linear power spectrum is generated at the mean pair redshift[1] of the [O II] ($z = 0.3$) and LAE ($z = 2.5$) samples using CAMB (Lewis et al. 2000). The modelling of the non-linear evolution of the power spectrum is based on a Galilean-invariant version of renormalized perturbation theory (Crocce & Scoccimarro 2006) dubbed gRPT, which will be presented in detail in Crocce et al. (in preparation) (see also the description in Eggemeier et al. 2020). The gRPT model offers a good description of the power spectrum down to $k \leq 0.25\,h^{-1}$ Mpc for a survey like BOSS (Sánchez et al. 2017).

The bias model we use for the input power spectrum is from Chan, Scoccimarro & Sheth (2012), and it relates the galaxy overdensity δ_g to the matter overdensity δ using local bias parameters consisting of b_1 and b_2 and non-local bias parameters γ_2 and γ_3^- as given in Chan et al. (2012). The full expression of the power spectrum from this bias model is given in appendix A of Sánchez et al. (2017). A review on perturbative bias is given in Desjacques, Jeong & Schmidt (2018)

To generate input power spectra for the mocks, the γ_2 and γ_3^- parameters are set following the local Lagrangian approximation (see Fry 1996; Catelan et al. 1998; Catelan, Porciani & Kamionkowski 2000; Chan et al. 2012). The local bias parameters we use for the LAEs and [O II] galaxies are $b_1 = 2.5$ and 1.5, respectively. The LAE bias we adopt is consistent with the $z \sim 2.5$ measurement of Khostovan et al. (2019), if we convert their power-law fits of the clustering to a bias via Quadri et al. (2007), which uses an expression from Peebles (1980). The [O II] galaxy bias is chosen to be consistent with previous work on HETDEX contamination (Grasshorn Gebhardt et al. 2019). For the LAE second-order bias we use fitting functions of b_1 versus b_2 from Lazeyras et al. (2016), which they derive using the separate universe approach of Wagner et al. (2015). We use the Lazeyras et al. (2016) results at redshifts slightly higher than the maximum redshift they test, but they see no evidence of redshift dependence in their relations in the range they do test, $0 < z < 2$. The fitting function yields $b_2 = 0.986$ for the LAEs. We do not use the same fitting function for the [O II] galaxies, as it gives a negative power spectrum at scales important to the simulation. This is likely due to an insufficient number of terms in the expansion; correcting this issue would require higher order bias terms in the expansion. We therefore set $b_2 = 0$ for these galaxies since it gives a reasonable power spectrum. We do not model any dependence of the clustering of sources on luminosity or other galaxy properties as this should not impact our conclusions.

[1]the mean over all pairs of $(z_1 + z_2)/2$, where z_1 and z_2 are the redshifts of each galaxy in the pair.

2.3 Lognormal simulations

To generate mock catalogues with our desired power spectrum we use the lognormal simulation code presented in Agrawal et al. (2017). A full explanation of the generation procedure is given in the above paper, but we include a brief summary here. The code uses an input power spectrum $P^G(k)$ to generate a 3D Gaussian field on a grid in k-space, $G(\boldsymbol{k})$. It also generates random phases for each grid point and then carries out a Fourier transform to generate $G(\boldsymbol{x})$, a realization of a Gaussian random field with the power spectrum $P^G(k)$. It then transforms this field to yield a field with a lognormal distribution, $\delta(\boldsymbol{x})$. The input power spectrum $P^G(k)$ is chosen in such a way that this resultant lognormal field will have the desired power spectrum $P(k)$. In this case our non-linear power spectrum is used for the matter density field, and our non-linear power spectrum with the added effects of bias is used for the galaxy density field. Each cell of the galaxy density field is randomly populated with galaxies. The number of galaxies assigned to a cell is drawn randomly from a Poisson distribution with a mean of $\bar{n}(1 + \delta(\boldsymbol{x}))V_{\rm cell}$, where $\bar{n}$ is the number density of galaxies and $V_{\rm cell}$ is the cell's volume. The code also assigns a velocity to every cell using the linearized continuity equation in Fourier space on the simulated matter density field, using linear growth rates from CAMB (Lewis et al. 2000). Mock galaxies are then assigned the velocity of their cell.

The cell size we use in the simulations is 2.2 h^{-1} Mpc for our LAE mocks. For the mocks of the [O II] galaxies, we use a minimum scale of 0.88 h^{-1} Mpc; the smaller size compensates for the fact [O II] emitters are projected on to larger scales by their misclassification as LAEs. We expect resolution effects on scales to occur at least as small as twice the cell size, and we will label this scale on our plots.

2.4 Adding an observer and the angular selection function

To convert the simulated galaxies into a catalogue, we place an observer at an appropriate position in simulation coordinates and compute the right ascension, declination and redshift to each mock galaxy from this observer's view point. The location of the observer, the simulation cube dimensions and the coordinate system are chosen in such a way to ensure the whole volume of a HETDEX field is contained within the simulation. We assume the two widely separated Spring and Fall fields are independent, and we also assume the density fields of LAEs and [O II] galaxies are independent (as in Addison et al. 2019 and Grasshorn Gebhardt et al. 2019, we ignore the small, inferred correlations from gravitational lensing). We therefore simulate each population with separate lognormal simulations.

The line-of-sight (LOS) direction between the observer and every galaxy is computed, and each galaxy's velocity is projected on to the galaxy's LOS direction. These LOS velocities are used to apply the offsets to the galaxy's 'observed' redshift, in order to model redshift space distortions (RSDs). In these mocks we do not consider additional effects from the virial motions of galaxies within groups

Table 1. A short summary of the important assumptions and input parameters for the mocks of an idealized HETDEX survey.

Cosmology – flat ΛCDM (Planck Collaboration VI 2020, with a small modification; see the notes)				
H	67.36 km s^{-1} Mpc^{-1}			
$\Omega_b h^2$	0.022 37			
$\Omega_c h^2$	0.12			
Ω_k	0			
n_s	0.9649			
σ_8	0.8226			
σ_{12}	0.8167			
LAE luminosity and EW functions (Gronwall et al. 2014)				
Redshift	2.063	3.104		
$L^*(h = 0.7)$(erg s^{-1})	4.07×10^{42}	5.98×10^{42}		
$\phi^*(h = 0.7)$(Mpc^{-3})	8.32×10^{-4}	1.05×10^{-3}		
α	−1.65	−1.65		
w_0 (Å)	50	100		
[O II] luminosity and EW function (Ciardullo et al. 2013)				
Redshift	0.1	0.2625	0.3875	0.5050
$L^*(h = 0.7)$(erg s^{-1})	1.17×10^{41}	1.95×10^{41}	3.16×10^{41}	3.79×10^{41}
$\phi^*(h = 0.7)$(Mpc^{-3})	5.01×10^{-3}	7.59×10^{-3}	8.51×10^{-3}	8.51×10^{-3}
α	−1.2	−1.2	−1.2	−1.2
w_0 (Å)	8.00	11.5	16.6	21.5
Survey properties (Sections 2.4 and 2.6)				
Field	Spring	Fall		
Total area (with gaps) (deg^2)	390	150		
Total area (covered by fibres) (deg^2)	55.6	27.2		
Volume with LAEs (h^{-3}Gpc^3)	2.42	0.93		
Number of IFUs	78	78		
Number of LAEs	6.4×10^5	2.9×10^5		
Number of [O II] galaxies	4.2×10^5	2.0×10^5		
LAE number density (h^3Mpc^{-3})	2.7×10^{-4}	3.1×10^{-4}		

Notes. The cosmological parameters are from Planck Collaboration VI (2020). The values for angular area of the survey are explained in Section 2.4, the prediction for the number of LAEs is explained in Section 2.6. The volume given is for the LAE redshift range and for the total area that is covered with gaps and sparse observations (see Chiang et al. 2013). The number density assumes the total number of LAEs are spread uniformly over that volume. As we, unlike Planck Collaboration VI (2020), assume massless neutrinos, we do not use their quoted σ_8 value but instead compute it using Lewis, Challinor & Lasenby (2000). We also include σ_{12}, the square root of the variance in 12-Mpc spheres (i.e. not using h units), as an alternative to the more standard σ_8 (see the arguments in Sánchez 2020).

and clusters or from Lyα radiative transfer (see e.g. Behrens et al. 2018; Byrohl, Saito & Behrens 2019; Byrohl et al. 2021; Gurung-López et al. 2019, 2020). Also note that although this modelling uses linear-theory-derived velocities, the resultant power spectrum in redshift space is subject to the non-linear aspects of RSD that arise from the transformation of mock galaxies from cosmological to observed redshifts (Agrawal et al. 2017).

We apply the angular footprints of the HETDEX fields to the mock catalogues. The exposure set positions for the full survey are combined with the expected positions of the full 78 IFUs in the focal plane. Instead of using the actual mask for the data taken on the telescope we use idealized exposure set positions and assume a full focal plane from the start. We also assume 78 working units for this analysis, as there remains a goal to reach this number on the telescope. Having 74 working units is a more realistic expectation given the data taken at the time of writing (Gebhardt et al. in preparation). These small differences should not impact our conclusions on the decontamination. Fig. 1 shows a mock catalogue with the angular selection function applied. The unusual shape of the Spring field is due to a decision (made in the first half of 2020) that the most efficient use of the telescope time is to extend the area rather than fill in missing regions from the originally planned footprint. This also explains the additional holes in the Spring footprint.

We use the masking software MANGLE from Hamilton & Tegmark (2004) and Swanson et al. (2008) to apply the survey footprint and also to generate a catalogue of random positions. These random positions are used to measure the clustering and we refer to them as the 'random catalogue' or 'randoms' hereafter. We also use MANGLE to compute the area of the sky covered by fibres: 55.6 deg^2 in the Spring field and 27.2 deg^2 in the Fall field, and make use of a wrapper to MANGLE called LITEMANGLE.[2] Although the footprint of HETDEX is unlikely to have any influence on our ability to discriminate LAEs from [O II] galaxies, it does influence the error estimates we use to assess the size of systematic biases.

These lognormal simulations are not true lightcone simulations like the ones used to probe contamination effects by Massara et al. (2020), or in the tomographic analysis of Awan & Gawiser (2020), since there is no evolution of the true power spectra along the redshift direction. The focus of this paper, however, is to determine

[2] https://github.com/martinjameswhite/litemangle.

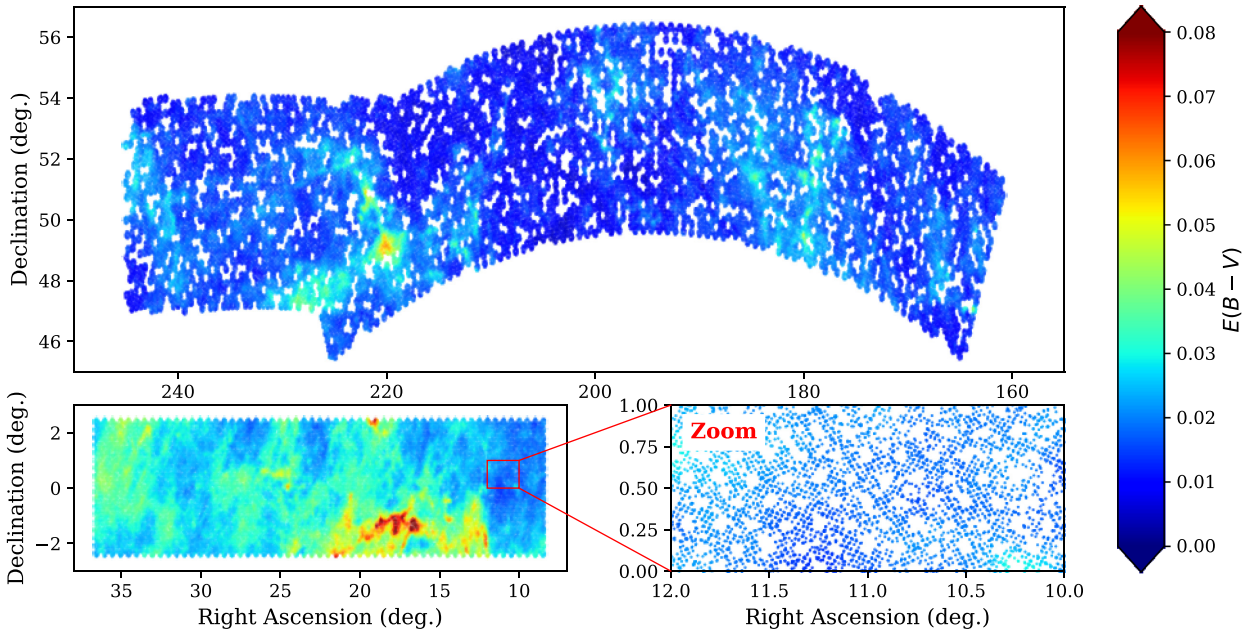

Figure 1. A scatter plot of the sources in one of our mock HETDEX survey Spring fields (top panels) and Fall fields (bottom left-hand panel), computed using an idealized focal plane containing 78 IFUs and a list of expected exposure set positions. The colour gives the reddening from Galactic extinction from the Schlegel, Finkbeiner & Davis (1998) dust maps. As the gaps between the IFUs are not visible on these two plots, we also show a zoom of the Fall field (bottom right-hand panel).

how the misclassification of [O II] emitters as LAEs produces a redshift dependent projection of the [O II] galaxy density field, and how this redshift dependence, combined with redshift-dependent contamination fractions, affects clustering. These effects are included as we compute 'observed' redshifts to all of our sources from a simulated observer's point of view.

2.5 LAE and [O II] properties

To generate catalogues with realistic number densities and classification probabilities, we need to assign points in our mocks luminosities and equivalent widths (EWs). To assign LAE luminosities we use the Schechter function fits to the $z = 2.1$ and 3.1 measured luminosity functions from Gronwall et al. (2014). These Schechter functions are parametrized by the characteristic luminosity, L^*, the faint-end slope, α, and the number density coefficient, ϕ^*. Similarly, we assume that the LAE EWs follow the exponential distributions found by Gronwall et al. (2014) at those two redshifts (see also equation 2 of Leung et al. 2017). The parameters for the [O II] luminosity and EW functions come from measurements in four redshift bins between $z = 0.1$ and 0.5 by Ciardullo et al. (2013). At redshifts other than bin centres, we use the linearly interpolated or extrapolated values of all of the parameters. In Table 1, we list the relevant parameters mentioned in this paragraph explicitly. The choice of luminosity and equivalent width functions are made to match the previous work on HETDEX source classification by Leung et al. (2017). We also use the approach of Leung et al. (2017) to correct the measured luminosity functions for low EW LAEs (EW < 20 Å), which were removed when the luminosity functions were estimated. We do not model any relationship between EW and luminosity as this level or realism is not needed for our work.

2.6 Assigning luminosities and the radial selection function

To apply the radial selection function to our mock and random catalogues, we first begin by assigning luminosities to our mock galaxies. A minimum luminosity is computed for each redshift assuming flux limits much deeper than those of the survey: 6×10^{-18} erg s^{-1} cm^{-2} for LAEs and 4×10^{-18} erg s^{-1} cm^{-2} for [O II] galaxies. A maximum luminosity is computed as a large multiple of the minimum value, $L_{\rm max} = 6000 L_{\rm min}$. To test if our choice of $L_{\rm max}$ could affect results, larger values were tested. To avoid having to run the full simulation pipeline, we adopted a faster approach to test where we integrated products of the mean extinction, the luminosity function and our completeness model in redshift slices up to an even higher $L_{\rm max}$. The number of sources predicted by our mocks and this simple integration-based technique agree to high precision (<1 per cent difference).

Between the two luminosity limits, random luminosities are drawn from our fiducial luminosity functions (Table 1). These luminosities are then translated to fluxes using the luminosity distance to the virtual observer. For the random catalogue, distances are randomly chosen in a way that is uniform in volume, and luminosities and fluxes are drawn that are consistent with that distance.

Model flux limits are adopted using the 5σ detection limit given in the HETDEX science requirements (and also presented in Hill et al. in preparation), and are based on a typical sky spectrum along with the project's expectations for image quality and the efficiency of the whole telescope, spectrograph and detector system. Achieving the number density of LAEs predicted by these flux limits is a target of HETDEX.

We divide the 5σ flux limit at each mock galaxy's observer-frame wavelength by 5, and use that value as the standard deviation of the Gaussian noise we add to the true flux of the emission line. The signal-to-noise (S/N) ratio of this noisy 'observed' emission line is computed and the line is classified as 'detected' in the simulation if

its observed S/N exceeds 5. This results in sources being detected 50 per cent of the time if their flux is exactly at the 5σ limit; the completeness that corresponds to other fluxes can be computed by integrating a Gaussian and determining the area above the S/N cut. Carrying out these mock observations is more resource intensive than simply applying the predicted $n(z)$ to the mocks, but it does have the benefit of including Eddington bias in the emission line fluxes. These fluxes are used to estimate the LAE/[O II] galaxy probabilities (see Section 2.10).

In addition to cuts in simulated S/N, another part of the radial selection function of [O II] emitters comes from their size. As explained by Leung et al. (2017), at $z < 0.05$ most [O II] emitters will appear extended in imaging data and therefore easily distinguished from the LAE sample. We therefore do not include $z < 0.05$ [O II] emitters in our simulations. The argument that imaging will be able to remove very nearby galaxies is also why we do not consider the impact of other even longer rest-frame wavelength potential contaminants such as [O III] emitters. We further discuss the possible impact of other contaminants on cosmology in an upcoming paper (Farrow et al. in preparation).

2.7 On-sky sensitivity variations and extinction

The sensitivity and the flux errors of HETDEX vary from exposure set to exposure set, IFU to IFU, and even fibre to fibre. These variations are likely to change how well we can classify LAEs as a function of sky position in the real survey. We do not model on-sky variations in the sensitivity of the survey, as in this work we focus on effects along the LOS, which are likely much more important, since the clustering of the contaminants evolves quickly due to projection effects (see Section 3.2).

Although we do not model the aforementioned on-sky sensitivity variations, we do add some sky-position-dependent effects as we attenuate the fluxes and mock spectra (see Section 2.9) by Galactic extinction. We model the sky-position dependence of Galactic extinction via the python library of Green (2018), utilizing the dust maps of Schlegel et al. (1998). To model the wavelength dependence we use the EXTINCTION[3] library with the Fitzpatrick (1999) function and the parameters advocated in the appendix of Schlafly et al. (2010). We confirm that our code can reproduce the extinction in the SDSS bands (Doi et al. 2010) predicted by Schlafly & Finkbeiner (2011) using Munari et al. (2005) stellar spectra to 2 per cent accuracy. This is more than sufficient for our mock catalogues. We also add this extinction to the randoms, so it is accounted for when measuring the clustering.

The Galactic dust reddening versus position from Schlegel et al. (1998) is indicated in Fig. 1; the equatorial Fall field typically has more Galactic extinction than the Spring field.

2.8 Source density versus redshift

In the following sections we continue our explanation of the mocks with how we assign mock continuum values and use them to classify galaxies as LAE or [O II]. Before this, let us consider the number density of the mock catalogues without contamination. Our model of the selection function predicts about a million LAEs and 600 000 $z > 0.05$ [O II] emitters in the full HETDEX survey. Fig. 2 shows the number density of detected emission line sources in one of our mock catalogues (solid lines) and in our random catalogue (dashed lines) versus redshift. The plots are computed using the full volumes

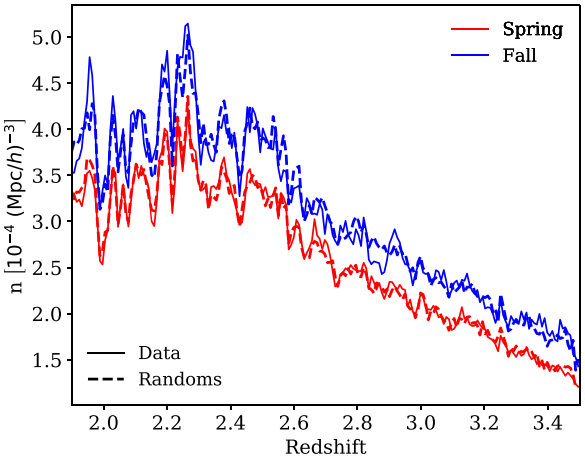

Figure 2. The number density of LAEs in one of our mock catalogues (solid lines) and random catalogues (dashed lines, normalized to the total number of mock sources) in the two HETDEX fields. The structure in the randoms is caused by the complex, wavelength-dependent flux limits. The number density is computed assuming the sparsely sampled on-sky area of the two fields; the Fall field has higher number density as the fill-factor of the area is larger.

of the two fields. The most prominent troughs in the number density of the randoms are not due to noise, but the effect of sky lines, which propagate into the survey's sensitivity limit. The difference in the number density between the Spring and Fall fields is mostly caused by different sky filling factors (i.e. there are more gaps in the Spring field). This figure shows the effect of the complicated radial selection function on the detected number density.

2.9 Mock LAE/[O II] galaxy spectra

In order to model the separation of LAE and [O II] emitters as accurately as possible, we generate mock spectra, which allow us to model the noise on the measured EWs more accurately. To do this we follow the approach of Leung et al. (2017). Details are available in that paper, but summarizing the method is helpful for future discussion. Equivalent widths are drawn from the distributions described in Section 2.5, with scale lengths as given in Table 1. A spectral slope is assigned to the line emitters, based on $(g - r)$ colours in SDSS filters (Doi et al. 2010) randomly selected from a distribution that looks like the real data (details in Leung et al. 2017). The line flux divided by the EW sets the amplitude of the mock spectra. Then absorption from the intergalactic medium is applied to the mock spectra from the prescription in Madau (1995), using the code adapted from Leung et al. (2017) and Acquaviva, Gawiser & Guaita (2011).

We apply broad-band filters to the mock spectra to simulate the imaging surveys we intend to use to make estimates of the continuum flux density. In the Fall field we already have Dark Energy Camera (DECam; Flaugher et al. 2015) *r*-band survey data from the *Spitzer*/HETDEX Exploratory Large Area survey (SHELA; Papovich et al. 2016; Wold et al. 2019), and the Dark Energy Survey (DES; Abbott et al. 2018), so we apply the DECam *r*-filter (Abbott et al. 2018). In the Spring field, we have complete coverage with Hyper-Suprime Cam (HSC) data in the *r*-band, so we apply the HSC filter (Kawanomoto et al. 2018). We use the PYTHON library

[3] https://extinction.readthedocs.io/en/latest/.

SPECLITE[4] to supply the filter response functions. Noise is added to the mock magnitude measurements, using a rough estimate derived by dividing the 5σ flux limits of the SHELA survey by five. The 5σ sky-aperture magnitude limits of SHELA were determined by Wold et al. (2019), and we take the mean of the four different fields in this work, $r = 24.6$, converting into flux via Oke & Gunn (1983). For simplicity, we use the noise based off of SHELA for the whole survey, which in some areas is actually covered by DES or HSC. This simplification has some impact on the precision of the assigned probabilities, but should not affect the conclusions of our work. Also, early analysis suggests the HSC data is significantly deeper than the SHELA data, so in the Spring field this is a conservative approach. The noisy magnitude measurements are combined with the noisy line flux measurements to make a noisy estimate of the equivalent width, $EW_{\rm obs}$.

A few subtleties are worth mentioning here. Firstly, although all of the noise we add is Gaussian, the distribution of $EW_{\rm obs}$ can be realistically non-Gaussian due to taking the inverse of the noisy continuum estimates. Secondly, note we make the assumption in our mock EW observations that the continuum is flat across the r-band and the spectral range of HETDEX. In real data more sophisticated techniques could be used, but here we again decide to be conservative and make the most simple mock measurements from our spectra. Finally, note that for the broad bands we use, which are to the red of Ly α, applying IGM absorption makes no difference to the results, but we include it in the model for possible future work.

Also following the Leung et al. (2017) approach, we add other expected emission lines to the spectra of [O II] galaxies (namely [Ne III] λ3869 Å, H β λ4959 Å, [O III] λ4949 Å, and [O III] λ5007 Å) using fixed line ratios for one fifth solar abundance (Anders & Fritze-v. Alvensleben 2003, and references therein). We also add appropriate Gaussian noise to these lines, following the same wavelength-dependent noise prediction used for Ly α. These other emission lines can also be used to identify [O II] emitters in the regions of redshift where they are within the spectral range of HETDEX. We use this method to generate 1000 realistic mock HETDEX catalogues.

2.10 A modified method to assign probabilities

To split the mocks into 'observed' LAE and [O II] samples, we assign each mock source a probability of being an LAE, based on its 'observed' properties. To generate these probabilities, we reformulate the Bayesian method of separating the two classes that was presented in Leung et al. (2017). We begin by presenting a conceptually different way to formulate the problem, that results in a set of more easily evaluated equations. We use the same set of inputs as in Leung et al. (2017), except for the source colour as it is unclear whether we will have deep multiband imaging over the whole HETDEX field. We then consider a small n-dimensional box in the parameter space of EW, flux, wavelength, and the flux of other non-[O II]/LAE emission lines. Assuming the primary emission -line can only be [O II] λ3727 or Ly α, the probability of the source being an LAE is

$$P_{\rm LAE} = \frac{N_{\rm LAE}}{N_{\rm LAE} + N_{\rm [O\,II]}}, \tag{1}$$

[4]Note we use the older 'DECam 2014' filters; see the SPECLITE website for details (https://speclite.readthedocs.io/en/latest/index.html). Using the older filter curves should not impact our conclusions.

where $N_{\rm LAE}$ and $N_{\rm [O\,II]}$ represent the number of LAE and [O II] emitters, respectively, in the box defined in the space of parameters used for the discrimination. We want this box to be a fixed size in observed coordinates. If we choose a fractional interval of $\pm\delta$ in observed flux (f), equivalent width, (w) and wavelength (λ) this corresponds to

$$(1 \pm \delta)L = (1 \pm \delta) f \times 4\pi d_{\rm L}^2, \tag{2}$$

$$(1 \pm \delta)w = (1 \pm \delta) w_{\rm obs}/(1 + z), \tag{3}$$

$$(1 \pm \delta)(z + 1) - 1 = (1 \pm \delta)\lambda/\lambda_{\rm line} - 1, \tag{4}$$

where $d_{\rm L}$ is the luminosity distance. We can now express the number in terms of integrals over the luminosity function, $\Phi(L/L_*, z)\,{\rm d}L/L^*$, the equivalent width distribution $W(w, z)$ and a Gaussian, $G(f_{\rm obs} - f_{\rm exp}, \sigma_{\rm line})$, with mean $f_{\rm exp}$ and dispersion $\sigma_{\rm line}$. This last term expresses the difference between the noisy measured flux and the expected flux, $f_{\rm exp}$, of a non-[O II] emission line (i.e. [Ne II], [O II] etc.), in terms of the uncertainty in the measurement, $\sigma_{\rm line}$. This term is the product over all of the other emission lines that are expected, given the wavelength of detection and assuming the galaxy is an [O II] emitter. The expression for the expected number of LAEs or [O II] galaxies is then

$$\begin{aligned} N = & \int_{(z+1)(1-\delta)-1}^{(z+1)(1+\delta)-1} \frac{{\rm d}V}{{\rm d}z}{\rm d}z' \int_{L(1-\delta)}^{L(1+\delta)} \Phi(L'/L_*, z){\rm d}(L'/L_*) \\ & \times \prod_{i=[{\rm O\,III}],[{\rm H}\beta],\ldots} \int_{f_{{\rm obs},i}(1-\delta)}^{f_{{\rm obs},i}(1+\delta)} G(f' - f_{{\rm exp},i}, \sigma_i){\rm d}f' \\ & \times \int_{w(1-\delta)}^{w(1+\delta)} W(w', z){\rm d}w'. \end{aligned} \tag{5}$$

This equation is very similar to equation (19) of Leung et al. (2017), except here we do not normalize by the number density of the emission line sources at the redshift under consideration. Moreover, Leung et al. (2017) chose a fixed size value for δ; we set δ to be infinitesimally small as then we can drop the integrals. The number in an infinitesimally sized box becomes

$$\begin{aligned} N = & \frac{{\rm d}V}{{\rm d}z} 2\delta(z + 1) \cdot \Phi(L'/L_*, z) 2\delta L/L_* \cdot W(w', z) 2\delta w \\ & \times \prod_{i=[{\rm O\,III}],[{\rm H}\beta],\ldots} G(f_{{\rm obs},i} - f_{{\rm exp},i}, \sigma_i) 2\delta f_{{\rm obs},i}. \end{aligned} \tag{6}$$

Then, using equation (1), substituting $1 + z$ with the ratio of observed to assumed rest wavelength, and cancelling the 2δ and λ terms, the expression for the LAE probability becomes

$$P_{\rm LAE} = \frac{\tilde{N}_{\rm LAE}}{\tilde{N}_{\rm LAE} + \tilde{N}_{\rm [O\,II]}}, \tag{7}$$

with

$$\begin{aligned} \tilde{N}_x = & \Lambda_x \frac{{\rm d}V}{{\rm d}z} \Phi(L_{x/L_{*,x}}, z_x) \frac{L_x}{L_{*,x}} W_x(w_x, z_x) w_x \\ & \times \prod_{i=[{\rm O\,III}],[{\rm H}\beta],\ldots} G(f_{{\rm obs},i} - f^x_{{\rm exp},i}, \sigma_i) f_{{\rm obs},i}, \end{aligned} \tag{8}$$

where x labels whether the relevant functions and measurements are for LAEs or [O II] galaxies, $\Lambda_{\rm LAE} = 1$ and $\Lambda_{\rm [O\,II]} = \lambda_{\rm LAE}/\lambda_{\rm [O\,II]}$. For LAEs, the expected flux at the wavelength of other emission lines is $f^{\rm LAE}_{{\rm exp},i} = 0$, while for [O II] emitters, this value is equal to the relative line ratio for each line, R_i, multiplied by the observed [O II] flux, i.e. $f^{\rm [O\,II]}_{{\rm exp},i} = R_i f_{\rm obs,[O\,II]}$. In these simulations we evaluate equation (8)

using the true underlying input luminosity and equivalent width distributions, the input line ratios, and cosmology used in the survey. We also use our mock observed measurements when computing the probabilities, which adds noise similar to real data. Future HETDEX papers will carry out more extensive tests and assessments of LAE classification approaches (Davis et al. in preparation).

Our library to produce these probabilities, and also an implementation of the Leung et al. (2017) method, has been integrated into the rest of the HETDEX source classification code, and is also available online.[5] The authors of Leung et al. (2017) provided us with their original code, which we use as a reference (and for some sections reproduce directly) in our implementation. This is also true for parts of the HETDEX simulation pipeline.

2.11 The mock observed LAE and [O II] samples

To generate samples of contaminated LAEs and [O II] emitters from the mocks, we classify all sources with $P_{\rm LAE} > 0.5$ as LAE and all other sources as [O II] galaxies. Despite the fact that these probabilities do not account for the noise on the $EW_{\rm obs}$ or on the LAE/[O II] line flux, this simple cut produces an LAE sample where only 1.3 per cent of the sources are misclassified [O II] emitters and 4.4 per cent of the observed [O II] catalogue are LAEs. This is actually better than the target LAE sample contamination fraction of 2 per cent, but our classifier is better than what is obtainable for real data, as it assumes we know the properties of the input LAE and [O II] populations perfectly. To consider a pessimistic scenario we also split the samples using a less conservative cut of $P_{\rm LAE} > 0.15$, which produces a purer [O II] sample (contamination fraction of 1.7 per cent), but a greater number of contaminants in the LAE sample (5.1 per cent). It might seem surprising that the $P_{\rm LAE} > 0.15$ cut still gives a relatively small fraction of contaminants, but it is important to realize the $P_{\rm LAE}$ values assigned to individual [O II] emitters are skewed towards zero, as for most sources, the classification is nearly unambiguous. In the rest of the paper we will refer to the *high-contamination* sample as that for $P_{\rm LAE} > 0.15$ and the *low-contamination* sample for $P_{\rm LAE} > 0.5$. These two samples bracket the expected 2 per cent contamination of HETDEX.

In order to create a random catalogue that correctly follows the redshift distribution of the data samples, we also compute LAE probabilities for the random catalogue and apply the same probability cuts. If we used random catalogues without contamination the different redshift distribution of the randoms versus that inferred for the observed samples would cause a huge systematic bias.

The predicted sample purity, defined as the number of correctly classified sources in a sample divided by the total size of the sample, is shown in Fig. 3. The lower redshift limit of this plot corresponds to our minimum redshift for [O II] emitters ($z = 0.05$). Although our simulations make the simplifying assumption of perfect knowledge of the true distribution of [O II] and LAE properties, we can still see many features expected for LAE/[O II] classifiers. As the observed emission line wavelength increases, the volume of space inhabited by [O II] emitters grows faster than that of the LAEs, causing a decrease in the purity of the LAE sample. The large, sudden decreases in the purity correspond to where emission lines useful in identifying a source as an [O II] emitter are redshifted out of the HETDEX spectral range. Although the full, high-contamination LAE sample has an interloper fraction of 5.1 per cent, when the sample is split by redshift, the contamination can be as large as around 17 per cent in the highest redshift bins.

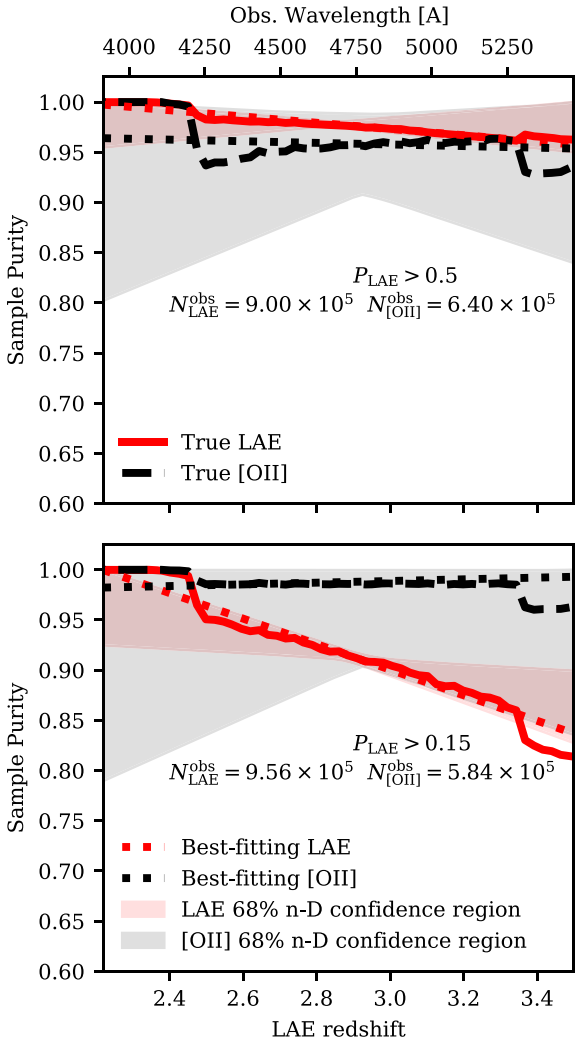

Figure 3. The purity of the mock 'observed' LAE (solid red lines) and [O II] (dashed black lines) galaxy catalogues for the $P_{\rm LAE} = 0.5$ (top panel) and 0.15 cuts (bottom panel). We only show the observed wavelength range where [O II] emitters are included in the simulation. The sharp drops occur where important emission lines redshift out of the HETDEX spectral range, specifically [O III] λ5007 [O III] λ4949, H β, and [Ne III] at $z = 2.35$, 2.38, 2.45, and 3.34, respectively. The inset numbers show the total number of sources in the full HETDEX redshift range in each of the samples (including interlopers) for the given cuts. The dotted lines show the best-fitting contamination values from our linear model of LAE and [O II] purity, which has two parameters per galaxy type: $f(z_{\rm low})$ and $f(z_{\rm high})$. The shaded regions show the maximum and minimum purity values in the 68 per cent confidence region (see Section 6.4).

3 CORRELATION FUNCTIONS

3.1 Measuring the clustering

The correlation functions of the mock catalogues are measured on a two-dimensional grid of the galaxy and/or random pair separation, s, and the cosine of the angle between the pair separation vector and the LOS, μ. We use the estimator introduced by Landy & Szalay (1993),

[5] https://github.com/djfarrow/hetdex-line-classification.

modified for cross-correlation functions by Blake et al. (2006),

$$\xi(s,\mu) = \frac{DD_{\rm c}(s,\mu) - D_{\rm c}R(s,\mu) - DR_{\rm c}(s,\mu) + RR_{\rm c}(s,\mu)}{RR_{\rm c}(s,\mu)}, \quad (9)$$

where c indicates which of the objects in the pair is an [O II] emitter and $DD_{\rm c}(s,\mu)$, $D_{\rm c}R(s,\mu)$, $DR_{\rm c}(s,\mu)$, and $RR_{\rm c}(s,\mu)$ are the binned counts of pairs of LAEs and [O II] galaxies, [O II] galaxies and LAE randoms, LAEs and [O II] randoms, and LAE randoms and [O II] randoms, respectively. The autocorrelation functions are estimated with the usual Landy & Szalay (1993) estimator. We compute the LOS direction to each pair of galaxies as the vector between the observer and the mid-point of the separation vector of the pair. When measuring the correlation function, we use random LAE and/or [O II] catalogues at least 13 times larger than the data catalogue, to decrease shot noise from the randoms. We measure the autocorrelation functions and the cross-correlation functions assuming Ly α derived redshifts for both the LAE and [O II] catalogues, except when measuring the [O II] clustering to use with equation 14, where we use the [O II] derived redshifts.

The 2D correlation functions are integrated in μ, weighted with the appropriate Legendre polynomials, to yield measurements of the first three even multipoles, $\xi_\ell(s)$, following the standard method (e.g. Sánchez et al. 2017). The covariance matrix is estimated from the measured multipoles also using the standard approach, i.e.

$$C_{\ell\ell'}(s_a, s_b) = \frac{1}{N_{\rm mk}-1}\sum_{i=0}^{N_{\rm mk}} \left(\xi_\ell(s_a) - \bar{\xi}_\ell(s_a)\right)\left(\xi_{\ell'}(s_b) - \bar{\xi}_{\ell'}(s_b)\right), \quad (10)$$

where $C_{\ell\ell'}(s_a, s_b)$ is the covariance between multipoles ℓ and ℓ', for measurement bins s_a and s_b. The index i runs over the number of mock catalogues, $N_{\rm mk} = 1000$. The quantities with bars, e.g. $\bar{\xi}_{\ell'}(x_b)$, are the mean values from all of the mock catalogues.

The simulated Fall and Spring fields have different average flux limits due to different values of the Galactic extinction. Normally, if the fields have significantly different average flux limits, they would be biased differently and need to be analysed separately. In our simulations however all the LAE sources have the same correlation function; we therefore combine the two fields by computing weighted sums of the multipoles and covariances following equations (8) and (9) of White et al. (2011).

3.2 Projected [O II] clustering

The [O II] contaminants in the LAE sample are assigned redshifts assuming the rest-frame wavelength of Ly α, and vice versa for the LAE contaminants in the [O II] sample. The relation between the source redshift assuming the emission line is [O II] λ3727 rather than Ly α is simply given by

$$z_{\rm [O\,II]} = (1 + z_{\rm LAE})\frac{\lambda_{\rm LAE}}{\lambda_{\rm [O\,II]}} - 1. \quad (11)$$

As noted in Lidz & Taylor (2016), the misclassification has an effect very analogous to the Alcock–Paczynski test (Alcock & Paczynski 1979, hereafter AP), in that the three-dimensional positions inferred from the position and redshift of the sources are distorted. Following Pullen et al. (2016), Leung et al. (2017), and the earlier similar derivation from Visbal & Loeb (2010) while adopting a slightly different notation, we can relate the true separation of a pair of [O II] emitters, in directions parallel, $s'_\parallel$, and perpendicular, $s'_\perp$, to the LOS, to the separation projected into LAE coordinates ($s_\perp$, $s_\parallel$) by

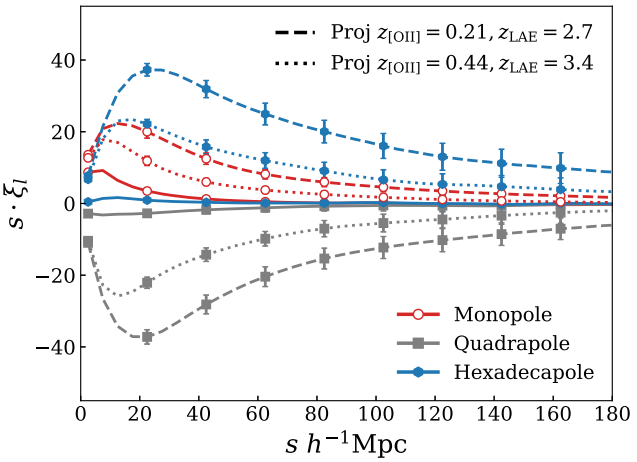

Figure 4. The solid lines show mean of the [O II] galaxy correlation function multipoles measured from 199 of our mock catalogues, along with error bars expected from a single realization of HETDEX. The dotted and dashed lines show the multipoles when distorted by a projection to different LAE redshifts, as indicated in the legend. This projection occurs due to LAE/[O II] misclassification and we model it using equation (14). For visual clarity, only every fourth data point and error bar is marked, and the correlation functions have been multiplied by the separation, s.

misclassification with

$$s'_\perp = s_\perp c_\perp, \qquad s'_\parallel = s_\parallel c_\parallel, \quad (12)$$

with

$$c_\parallel(z_{\rm LAE}) = \frac{\lambda_{\rm LAE}}{\lambda_{\rm [O\,II]}}\frac{H(z_{\rm LAE})}{H(z_{\rm [O\,II]})}, \qquad c_\perp(z_{\rm LAE}) = \frac{D_{\rm M}(z_{\rm [O\,II]})}{D_{\rm M}(z_{\rm LAE})}, \quad (13)$$

where $H(z)$ is the Hubble parameter and $D_{\rm M}(z)$ is the comoving angular diameter distance to z. This is given by $D_{\rm M}(z) = (1 + z)D_{\rm A}(z)$, where $D_{\rm A}(z)$ is the angular diameter distance. Given these distortion parameters, the correlation function can be written as

$$\xi^{\rm proj}_{\rm [O\,II]}(s,\mu,z_{\rm LAE}) = \xi_{\rm [O\,II]}\left(sq(\mu), \mu c_\parallel(z_{\rm LAE})/q(\mu)\right), \quad (14)$$

where we do not explicitly show the dependence of q on $z_{\rm LAE}$ to shorten the notation (for the expression for the power spectrum see e.g. Pullen et al. 2016; Leung et al. 2017; Grasshorn Gebhardt et al. 2019). Equation (14) assumes all the evolution of the projected [O II] clustering is caused by projection effects, as is the case in our simulations. It should be possible in future work to extend this methodology to also include intrinsic evolution of the [O II] correlation function. The value of q is given by Ballinger, Peacock & Heavens (1996) (see also e.g. equation 9 of Pullen et al. 2016):

$$q(\mu) = \left[c_\parallel^2(z_{\rm LAE})(\mu)^2 + c_\perp^2(z_{\rm LAE})(1-(\mu)^2)\right]^{1/2}. \quad (15)$$

The equations describing the clustering of LAEs misclassified as [O II] galaxies are the same but with the inverse of the distortion parameters, i.e. $c_\parallel^{-1}$ and $c_\perp^{-1}$. As the distortion parameters are an approximation of a more complicated effect, we carry out tests in Appendix A of the distortion parameters compared to a brute force approach. This appendix also presents an additional test of the methodology we present in Section 4.2.

The redshift dependence of the distortions causes the clustering of the [O II] contaminants to evolve with (Ly α-based) redshift. To illustrate these effects we show in Fig. 4 the mean correlation function measured from 199 pure mock [O II] catalogues, along with the same measurements projected on to two different Ly α redshifts. To

predict the projected measurements, we use equation (14), linearly interpolating over the measured [O II] correlation function for $\xi_{[\mathrm{O\,II}]}$. The solid lines show multipoles from the samples analysed with the [O II] redshifts, the negative quadrapole is evidence of the Kaiser effect (Kaiser 1987), an effect of the peculiar velocities of galaxies falling into overdensities. The dashed line shows the predictions of projecting from the [O II] redshift at $z_{[\mathrm{O\,II}]} = 0.21$ to the misclassified LAE redshift of $z_{\mathrm{LAE}} = 2.7$. We see the projection causes a clear increase in the monopole for all but the smallest separations under consideration. We also see the quadrupole becomes much more negative, which is a result of the projected correlation function appearing very elongated along the direction transverse to the LOS. The impact of this on the multipoles is much larger than for the Kaiser effect. In Fourier space, the elongation looks like a compression along the direction transverse to the LOS (see e.g. fig. 3 of Grasshorn Gebhardt et al. 2019). We also note an increase in the hexadecapole.

The dotted lines in Fig. 4 show the predicted multipoles of $z_{[\mathrm{O\,II}]} = 0.44$ [O II] emitters that are misclassified as $z_{\mathrm{LAE}} = 3.4$ LAEs. We see similar trends to the lower redshift projection, but the amplitude of the distorted multipoles is lower. This decrease is driven by $c_\perp$ becoming closer to unity, and the distortion transverse to the LOS is much larger than the distortion in the parallel direction, $c_\parallel$ (e.g. Grasshorn Gebhardt et al. 2019). We will return to modelling these signals over a redshift range in Section 4.2.

At this point, we highlight the fact that we always use the true cosmology when computing the parameters for the projection. As highlighted by Addison et al. (2019), if we want to make predictions for the projected functions in the real data, we need to be aware of the additional uncertainty from not knowing the actual cosmology. We discuss this again at the end of this paper.

4 DECONTAMINATION METHODS

4.1 Simple decontamination ignoring redshift dependencies

As mentioned, in this paper we develop a new method to deal with the redshift dependence of the contamination. We start by slightly modifying equation (12) of Awan & Gawiser (2020) to use the multipoles of the two-dimensional correlation function instead of the angular clustering, giving

$$\left[\xi^{\mathrm{obs}}_{\ell,\mathrm{aa}}(s), \xi^{\mathrm{obs}}_{\ell,\mathrm{ab}}(s), \xi^{\mathrm{obs}}_{\ell,\mathrm{bb}}(s)\right]^{\mathrm{T}} = \mathbf{D}_{\mathrm{s}} \left[\xi^{\mathrm{true}}_{\ell,\mathrm{aa}}(s), \xi^{\mathrm{true}}_{\ell,\mathrm{ab}}(s), \xi^{\mathrm{true}}_{\ell,\mathrm{bb}}(s)\right]^{\mathrm{T}}, \quad (16)$$

where ℓ indicates the multipole, 'a' and 'b' indicate the two possible samples (in our case LAEs and [O II] galaxies), the 'true' and 'obs' superscripts indicate the pure and contaminated correlation functions and $\mathbf{D}_{\mathrm{s}}$ is the contamination matrix. The matrix of Awan & Gawiser (2020) compactly expresses the important equations for contamination, which have also been presented in other literature (e.g. Pullen et al. 2016; Leung et al. 2017; Addison et al. 2019; Grasshorn Gebhardt et al. 2019). The matrix contains contributions from the fractions of each type of galaxy that were correctly classified (i.e. the purity), labelled f_{aa}, and f_{bb}, and the fractions that were misclassified, f_{ab} and f_{ba}. In Awan & Gawiser (2020), this matrix is given as

$$\mathbf{D}_{\mathrm{s}} = \begin{pmatrix} f_{\mathrm{aa}}^2 & 2 f_{\mathrm{aa}} f_{\mathrm{ab}} & f_{\mathrm{ab}}^2 \\ f_{\mathrm{aa}} f_{\mathrm{ba}} & f_{\mathrm{aa}} f_{\mathrm{bb}} + f_{\mathrm{ab}} f_{\mathrm{ba}} & f_{\mathrm{ab}} f_{\mathrm{bb}} \\ f_{\mathrm{ba}}^2 & 2 f_{\mathrm{bb}} f_{\mathrm{ba}} & f_{\mathrm{bb}}^2 \end{pmatrix}, \quad (17)$$

where the contamination fractions can be computed from the purity via $f_{\mathrm{ba}} = 1 - f_{\mathrm{bb}}$ and $f_{\mathrm{ab}} = 1 - f_{\mathrm{aa}}$. To be more specific to the case of HETDEX, we relabel f_{aa} as f_{LAE} and f_{bb} as $f_{[\mathrm{O\,II}]}$. Also following Awan & Gawiser (2020), the decontaminated estimates of the auto and cross-correlation functions can then be given by applying the matrix inverse to a vector of the observed functions, i.e.

$$\left[\xi^{\mathrm{est}}_{\ell,\mathrm{aa}}(s), \xi^{\mathrm{est}}_{\ell,\mathrm{ab}}(s), \xi^{\mathrm{est}}_{\ell,\mathrm{bb}}(s)\right]^{\mathrm{T}} = \mathbf{D}_{\mathrm{s}}^{-1} \left[\xi^{\mathrm{obs}}_{\ell,\mathrm{aa}}(s), \xi^{\mathrm{obs}}_{\ell,\mathrm{ab}}(s), \xi^{\mathrm{obs}}_{\ell,\mathrm{bb}}(s)\right]^{\mathrm{T}}. \quad (18)$$

The superscript 'est' indicates the decontaminated estimates of the correlation function. Again, for the specific case of HETDEX $\xi_{\ell,\mathrm{aa}}(s)$, $\xi_{\ell,\mathrm{bb}}(s)$ and $\xi_{\ell,\mathrm{ab}}(s)$ are the autocorrelation functions of the LAE sample, $\xi_{\ell,\mathrm{LAE}}(s)$, the [O II] sample, $\xi_{\ell,[\mathrm{O\,II}]}(s)$, and the cross-correlation $\xi_{\ell,\mathrm{LAE}\times[\mathrm{O\,II}]}(s)$ respectively. Once we have estimates of the autocorrelation functions, we can make an estimate for the contribution of the contamination to the observed cross-correlation signal, $\xi^{\mathrm{pred,obs}}_{\ell,\mathrm{LAE}\times[\mathrm{O\,II}]}(s)$, using equation (16), resulting in

$$\begin{aligned}\xi^{\mathrm{pred,obs}}_{\ell,\mathrm{LAE}\times[\mathrm{O\,II}]}(s) = {} & f_{\mathrm{LAE}}(1 - f_{[\mathrm{O\,II}]})\xi^{\mathrm{est}}_{\ell,\mathrm{LAE}}(s) \\ & + f_{[\mathrm{O\,II}]}(1 - f_{\mathrm{LAE}})\xi^{\mathrm{est}}_{\ell,[\mathrm{O\,II}]}(s). \end{aligned} \quad (19)$$

This can be related to the full decontaminated cross-correlation from equations (16) and (17) via

$$\xi^{\mathrm{est}}_{\ell,\mathrm{LAE}\times[\mathrm{O\,II}]}(s) = \frac{\xi^{\mathrm{obs}}_{\ell,\mathrm{LAE}\times[\mathrm{O\,II}]}(s) - \xi^{\mathrm{pred,obs}}_{\ell,\mathrm{LAE}\times[\mathrm{O\,II}]}(s)}{f_{[\mathrm{O\,II}]} f_{\mathrm{LAE}} + (1 - f_{[\mathrm{O\,II}]})(1 - f_{\mathrm{LAE}})}. \quad (20)$$

We will label this approach 'simple decontamination' and differentiate it from our new approach of 'lightcone decontamination'. We note that Awan & Gawiser (2020) developed this method for angular clustering in tomographic redshift bins. They do not claim that the method will work for our scenario, which has rapidly evolving projected [O II] contamination within the redshift bins considered. However, we present it in its unmodified form as a demonstration of what might happen if one does not take additional steps to deal with this rapid evolution.

4.2 Lightcone-based decontamination

When we apply the matrix of Awan & Gawiser (2020) to HETDEX, we make the assumption that the clustering of the galaxies classified as [O II] emitters is the same as the clustering of [O II] interlopers in the LAE sample with some fixed scaling for contamination. However, this may not be the case, as the shape of the volume number density versus redshift, $n(z)$, of the interlopers will not match that of the [O II] sample when the purity has a redshift dependence. To give a hypothetical example, consider most of the [O II] emitters being at the high redshift end of the range. If that were the case, the projected clustering of the [O II] sample would have distortion parameters appropriate for high redshifts. If all of the misclassifications occurred at low redshift, however, then the interlopers would have low-redshift distortion parameters.

The idea then is to use something like the decontamination matrix of Awan & Gawiser (2020), but instead of using the observed clustering of the [O II] emitters, we apply a prediction for the clustering of contaminants that is consistent with the redshift dependence of the interloper number density, $n^{\mathrm{inter}}(z)$. To make a prediction for the expected interloper clustering in a redshift range, we refer to the work of Yamamoto & Suto (1999) and Suto et al. (2000). They approximate the correlation function between two galaxies as the correlation function at the mid-point between them (equation 19 of Yamamoto & Suto 1999). This results in a fairly intuitive expression that approximates the observed correlation function for galaxies in the redshift range z_{min} to z_{max} as an integral of the redshift-dependent correlation function weighted by the square of the number density as

a function of redshift, i.e.

$$\xi_\ell^{LC}(s) = \frac{\int_{z_{\rm min}}^{z_{\rm max}} {\rm d}z \frac{{\rm d}V}{{\rm d}z} n(z)^2 \xi_\ell(s;z)}{\int_{z_{\rm min}}^{z_{\rm max}} {\rm d}z \frac{{\rm d}V}{{\rm d}z} n(z)^2}. \quad (21)$$

This differs slightly from equation (18) of Suto et al. (2000) in that we use the observed number density of objects, not the true comoving number density in real space, so that the terms related to the selection function and the AP distortion are unneeded. Equation (21) also assumes that $n(z)$ does not change much over the separations under consideration, $s < 180\,h^{-1}$ Mpc, and the redshift evolution of $\xi_\ell(s;z)$ is slow enough to be unimportant over those same scales. This is an approximation, as there are certainly redshifts over which the projected [O II] clustering changes rapidly. But as we will see, the simplification works reasonably well for our simulations. For surveys whose properties differ from those of mock HETDEX, it would be prudent to test the technique with tailored simulations.

To continue, we define a function to carry out the lightcone (i.e. redshift) integral, $\mathcal{F}(x, y)$, as

$$\mathcal{F}(x(z), y(z)) = \frac{\int_{z_{\rm min}}^{z_{\rm max}} {\rm d}z \frac{{\rm d}V}{{\rm d}z} x(z)^2 y(z)}{\int_{z_{\rm min}}^{z_{\rm max}} {\rm d}z \frac{{\rm d}V}{{\rm d}z} x(z)^2}. \quad (22)$$

Given equation (22), and using equations (16) and (17) with equation (21), we obtain the following expression for the prediction of the observed autocorrelation function for LAEs with redshift z and purity $f(z)$:

$$\begin{aligned}\xi^{\rm LC,obs}(s,\mu) = {} & \mathcal{F}\left[n_{\rm LAE}(z),\ f^2(z)\xi^{\rm true}_{\rm LAE}(s,\mu,z)\right] \\ & + \mathcal{F}\left[n_{\rm LAE}(z),\ (1-f(z))^2\xi^{\rm proj}_{\rm [O\,II]}(s,\mu,z)\right] \\ & + 2\mathcal{F}\left[n_{\rm LAE}(z),\ f(z)(1-f(z))\xi^{\rm true,proj}_{\rm LAE\times[O\,II]}(s,\mu,z)\right]. \end{aligned} \quad (23)$$

The subscripts on $n(z)$ indicate which observed sample redshifts versus volume number density should be used. To be closer to the numerical implementation we replaced the multipoles of equation (16) with the 2D correlation function; this makes no practical difference as the decontamination has no μ dependence so the order of decontamination and converting the measurements into multipoles is unimportant. The number densities are for the total 'observed' samples with contaminants. Here we use the fact that the (LAE) redshift distribution of the [O II] interlopers is given by $n^{\rm inter.}_{\rm [O\,II]}(z) = (1-f(z))n_{\rm LAE}(z)$. We explicitly include the redshift dependence of the purity parameters, to differentiate them from their redshift-independent versions: $f_{\rm LAE}$ and $f_{\rm [O\,II]}$. For brevity, we also drop the LAE subscript from the redshift dependent purity parameter in this section. Since the integral of the number density gives the total number of objects, the redshift-dependent and independent types of purity parameter are related by the lightcone integral, i.e. for the LAE sample:

$$f_{\rm LAE} = \mathcal{F}\left(\sqrt{n_{\rm LAE}(z)},\ f(z)\right), \quad (24)$$

and the equivalent for the [O II] sample.

Samples of emission-line galaxies with nearby rest-frame wavelengths, such as H β and [O III], will have a non-zero cross-correlation due to large-scale structure, and even distant samples like our LAE and [O II] galaxies will have a slightly non-zero cross-correlation due to cosmic magnification of the background galaxies by the foreground galaxies. As mentioned, our simulations do not include such magnification, as it is a very small signal. We therefore will now assume that the true cross-correlation between the [O II] and LAE samples is zero. This limits the method to scenarios, like HETDEX, where the contaminants are not correlated with the main sample. However, future work on surveys with sample-contaminant cross-correlations could still use equation (23) in an approach that tries to forward-model the relevant auto and cross-correlations.

Given that the cross-correlation is zero, to find our estimate for $\xi^{\rm true}_{\rm LAE}(s,\mu,z)$ we now make the assumption that the LAE clustering does not evolve with redshift, which allows us to take it out of the lightcone integral and we are left with

$$\begin{aligned}\xi^{\rm est}_{\rm LAE}(s,\mu) = {} & \mathcal{F}\left[n_{\rm LAE}(z),\ f^2(z)\right]^{-1}\left\{\xi^{\rm LC,obs}(s,\mu)\right. \\ & \left. - \mathcal{F}\left[n_{\rm LAE}(z),\ (1-f(z))^2\xi^{\rm proj}_{\rm [O\,II]}(s,\mu,z)\right]\right\}. \end{aligned} \quad (25)$$

The integrals over redshift are then all carried out numerically and an estimate of the true correlation function, $\xi^{\rm est}_{\rm LAE}$, can be made. Even when the assumption that $\xi^{\rm true}_{\rm LAE}(s,\mu,z)$ does not evolve over the whole survey is unreasonable, this approach can be utilized to estimate the observed correlation function in bins of redshift over which that evolution is expected to be small enough that this average provides a meaningful observable. Additionally, we discuss plans to relax this assumption in Section 7.

Given our estimate of $\xi^{\rm est}_{\rm LAE}(s,\mu)$, we can use the cross-correlation term of equations (16) and (17) to predict the observed cross-correlation measured with LAE redshifts as follows:

$$\begin{aligned}&\xi^{\rm LC,obs}_{\rm LAE\times[O\,II]}(s,\mu) \\ &= \mathcal{F}\left[\left\{n_{\rm LAE}(z)n^{\rm proj}_{\rm [O\,II]}(z)\right\}^{0.5},\ f(z)(1-f_{\rm [O\,II]}(z))\xi^{\rm est}_{\rm LAE}(s,\mu)\right] \\ &\quad + \mathcal{F}\left[\left\{n_{\rm LAE}(z)n^{\rm proj}_{\rm [O\,II]}(z)\right\}^{0.5},\ f_{\rm [O\,II]}(z)(1-f(z))\xi^{\rm proj}_{\rm [O\,II]}(s,\mu,z)\right] \\ &\quad + \mathcal{F}\left[\left\{n_{\rm LAE}(z)n^{\rm proj}_{\rm [O\,II]}(z)\right\}^{0.5},\right. \\ &\quad \left.\times\ \left\{f(z)f_{\rm [O\,II]}(z) + (1-f(z))(1-f_{\rm [O\,II]}(z))\right\}\xi^{\rm true,proj}_{\rm LAE\times[O\,II]}(s,\mu,z)\right], \end{aligned} \quad (26)$$

where we have restored the cross-correlation term, $\xi^{\rm true,proj}_{\rm LAE\times[O\,II]}(s,\mu,z)$. Here $n^{\rm proj}_{\rm [O\,II]}(z)$ is the redshift distribution of [O II] emitters projected into LAE redshifts. This latter relation can easily be measured by computing redshifts assuming Ly α for the galaxies in the [O II] sample. We cannot estimate the redshift dependent cross-correlation by simply re-arranging this expression, but if the cross-correlation is expected to be non-zero a forward modelling approach could be used. Here we forward-model the expected cross-correlation signal, using the fact $\xi^{\rm true,proj}_{\rm LAE\times[O\,II]}(s,\mu,z) = 0$ in HETDEX, which gives

$$\begin{aligned}&\xi^{\rm pred,obs}_{\rm LAE\times[O\,II]}(s,\mu) \\ &= \mathcal{F}[\{n_{\rm LAE}(z)n^{\rm proj}_{\rm [O\,II]}(z)\}^{0.5},\ f(z)(1-f_{\rm [O\,II]}(z))\xi^{\rm est}_{\rm LAE}(s,\mu)] \\ &\quad + \mathcal{F}[\{n_{\rm LAE}(z)n^{\rm proj}_{\rm [O\,II]}(z)\}^{0.5}, \\ &\quad f_{\rm [O\,II]}(z)(1-f(z))\xi^{\rm proj}_{\rm [O\,II]}(s,\mu,z)]. \end{aligned} \quad (27)$$

This equation is the lightcone version of equation (19). The use of the lightcone equations requires a model of the projected [O II] clustering. As in Section 3.2, we interpolate over the measured clustering of the [O II] sample, and then apply a redshift dependent projection via equation (14). To make a fairer comparison of decontamination techniques, we only interpolate over the measurement of the [O II] clustering for each single realization of the catalogue, and we use the observed [O II] catalogue, not the pure one. As the observed [O II] clustering has contamination, we experimented with an iterative process where we first decontaminate the [O II] clustering with the projected LAE correlation function, and then

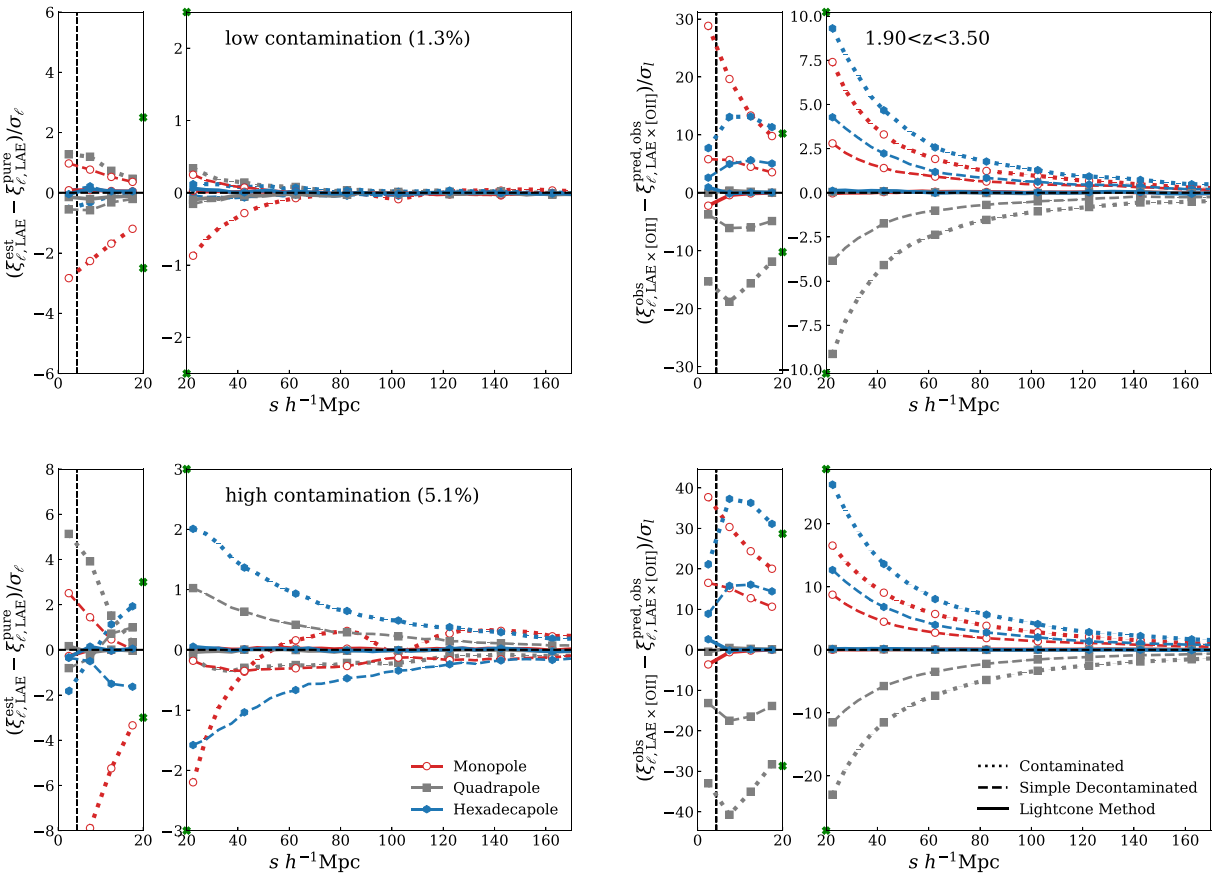

Figure 5. The left-hand column shows the average difference between the LAE correlation function from 1000 mock LAE catalogues containing [O II] contamination and the pure catalogues. Each panel is split at $s = 20\,h^{-1}$ Mpc to enable a better vertical and horizontal dynamic range, green crosses indicate the range of the right-hand panel on the axes of the left. In the right-hand panel, only every fourth data point is marked with a symbol for clarity. The right-hand column shows the residual signal left when subtracting the predicted cross-correlation from contamination from the measured cross-correlation functions, averaged over the 1000 mock [O II] and LAE catalogues. Different approaches to dealing with contamination are shown: dotted lines display results with no corrections, dashed lines show the results of a decontamination procedure that ignores the redshift dependence of the interlopers ('simple decontaminated') and the solid lines show results from our new method that accounts for redshift dependencies ('lightcone decontaminated'). If the contamination is fully accounted for, all the differences plotted here should be zero. All results are divided by the error on a single realization, to give a rough estimate of the statistical significance of any unaccounted for contamination in either method. The colours indicate the mono-(red), quadru- (grey), and hexadeca- (blue) poles of the correlation function. The different rows show results from the different LAE probability cuts used to define samples: 1.3 per cent contamination (top panels) and 5.1 per cent contamination (bottom panels); see Fig. 3 for the purity versus redshift of these samples. The dashed, black vertical line shows twice the cell size of the LAE simulation box.

use the decontaminated [O II] clustering to decontaminate the LAE clustering. We find the first and second iterations give almost identical results, so we stop after two iterations and use the resultant [O II] clustering to decontaminate the LAE measurement.

5 RESULTS

The differences between the multipoles of the autocorrelation function for the (de)contaminated and the pure cases (measured from the corresponding pure LAE catalogues) for the full HETDEX redshift range ($1.9 < z < 3.5$) are given in the left-hand column of Fig. 5. The upper and lower panels in Fig. 5 give the two $P_{\rm LAE}$ cuts under consideration. The points are the mean of the 1000 mocks. Each measurement has been divided by the statistical error expected for the HETDEX survey, i.e. the square root of the diagonal of the covariance matrix derived from the mocks. As we have 1000 mock catalogues, the errors on our mean measurement are much smaller than the statistical error on a single HETDEX mock. In the following, when we refer to σ, we specifically mean the statistical error on a single realization.

The right-hand column shows for the same redshift range, the residual difference between the observed cross-correlations of the [O II] and LAE samples, and the predicted cross-correlation from the contamination, from equation (19) for the simple method and equation (27) for the lightcone method. The differences are divided by the statistical errors. Recall from equation (20) that for the simple method the decontaminated cross-correlation is related to the residual we plot, $\xi^{\rm obs}_{\ell,\rm LAE\times[O\,II]}(s) - \xi^{\rm pred,obs}_{\ell,\rm LAE\times[O\,II]}(s)$, by a constant factor. As we divide by the statistical error, this means the residuals we plot for the simple decontamination method are also the statistical significance (ignoring the diagonal terms of the covariance) of the spurious cross-correlation signal that remains in the simple decontaminated multipoles. We chose to frame the discussion of the simple method in terms of the residuals, in order to make comparisons to the lightcone approach easier.

In the following subsections, we study the impact of contamination on the raw (Section 5.1), the simple decontaminated (Section 5.2),

and lightcone decontaminated (Section 5.3) multipoles. The final subsection considers the scenario where the redshift range with no modelled [O II] emitters ($z < 0.05$) is cut out of the catalogue (Section 5.4).

5.1 Raw clustering multipoles

It is clear from the dotted lines Fig. 5, which show the results from the raw measurements of the contaminated mocks, that interloper contamination modifies the correlation signals. This has been previously seen or predicted many times in the literature (e.g. Leung et al. 2017; Addison et al. 2019; Grasshorn Gebhardt et al. 2019; Awan & Gawiser 2020; Massara et al. 2020). In the low-contamination LAE sample, the offset in the autocorrelation is generally $<1\sigma$ but reaches a maximum difference of $\sim 2.5\sigma$ at scales down to twice the cell size of the lognormal simulations (indicated by the vertical dashed line). In the high-contamination sample, where 5.1 per cent of the LAE catalogue are [O II] galaxies, many of the autocorrelation function hexadecapole measurements are systematically high by up to 2σ at separations $>20\ h^{-1}$Mpc. At smaller scales, where the effect of the contamination appears greater, the raw autocorrelation function multipole measurements can be biased by several σ.

The signal of contamination in the cross-correlation function is even stronger. We can see from the right-hand column of Fig. 5 that we expect the signal to be clearly detected in the HETDEX survey, even at low levels of contamination. The cross-correlation function, which is zero for pure samples, shows a positive monopole and hexadecapole, and a negative quadrupole. The negative quadrupole implies that the two-dimensional cross-correlation function appears elongated transverse to the LOS. This could be explained by the fact that the cross-correlation is dominated by the strong clustering signal of [O II] galaxies (Grasshorn Gebhardt et al. 2019), which, as mentioned, appears elongated due to projection effects.

5.2 Simple decontaminated measurements

In this section, we use equation (18) to decontaminate the galaxy samples while ignoring redshift dependencies in the interloper fraction. We take the required contamination and purity factors, i.e. the components of $\mathbf{D}_s$, from the numbers of LAEs and [O II] emitters in the mock catalogues averaged over the realizations; in real data, some procedure will be needed to measure the contamination and purity. We cover this scenario in Section 6.

The dashed lines in Fig. 5, labelled 'simple decontaminated', show the results. In the low-contamination LAE sample (1.3 per cent [O II] galaxies), the simple method results in all of the multipoles having no significant systematic bias ($<1.0\sigma$) all the way down to the resolution limit of the catalogue. The decontaminated measurements are a modest improvement over the raw measurements. In the high LAE sample contamination case (5.1 per cent [O II] emitters), the decontamination improves the monopole, but at small scales, the monopole still has a bias approaching $\sim 2\sigma$. Additionally the hexadecapole is biased up to $\sim 1.5\sigma$ low after the decontamination – a bias in the opposite direction from what was seen in the raw data.

Subtracting the predicted contribution to the cross-correlation from contamination decreases the observed cross-correlation. However, the correction is too small, and very significant ($>5\sigma$ over a range of scales) cross-correlations still remain in the high-contamination case. The low-contamination sample also shows residual cross-correlation signals that increase relative to the statistical errors. As smaller separations are considered, these signals increase, causing what can be as great as an $\sim 5\sigma$ spurious signal. As mentioned, the residuals we plot are related to the full decontaminated cross-correlation from the simple method via a constant factor (equation 20), meaning they also show the significance of spurious cross-correlations left after the full decontamination process. Although a strong cross-correlation signal would not be expected to directly affect cosmological parameters derived from HETDEX, a non-zero cross-correlation after simple decontamination does suggest a failure of the modelling, which can cause indirect effects. For example, the methods presented in Addison et al. (2019) and Grasshorn Gebhardt et al. (2019) use the cross-correlation signal to determine the purity and contamination of the LAE and [O II] samples. If the lightcone effects are ignored, then inferred values of the contamination will become artificially high in order to force the decontaminated cross-correlation towards zero. This, in turn, would impact the contamination and purity values used to decontaminate the autocorrelation, causing additional bias in their measurement. We will demonstrate this in Section 6.

5.3 Lightcone decontamination

In this section we apply our new method of decontaminating the samples while accounting for lightcone/redshift effects. As with the simple method, we will start by assuming perfect knowledge of the purity and contamination, i.e. $f(z)$ and $f_{[\mathrm{O\,II}]}(z)$, and use the redshift-dependent contamination fraction measured from the random catalogues (i.e. Fig. 3). The results of this are the solid lines in Fig. 5.

In the low-contamination LAE sample autocorrelation function, we see little meaningful difference when compared to the simple method. Both return clustering multipoles with very little evidence of systematic bias. On the other hand, in the high-contamination sample, the lightcone-based contamination does a better job at correcting the multipoles. Down to a scale of $20\ h^{-1}$Mpc, the new method returns measurements with a bias less than $\sim 0.25\sigma$.

The differences between the lightcone model predicted and measured cross-correlation functions show an even greater improvement over the simple method. In both the low- and high-contamination scenarios, the new method accounts for the spurious cross-correlation signal leaving less than $\sim 1\sigma$ residuals at all scales greater than twice the cell size of the LAE simulation box.

The improvements seen in our approach support the idea that the residual, biased signals seen when using simple decontamination come from applying it to clustering measurements without accounting for the significant redshift evolution of the projected clustering and the contamination fraction within the redshift bin.

5.4 Restricted redshift range

The redshift range studied so far, $1.9 < z < 3.5$, includes a volume in which we expect there to be no [O II] emitters. This is strictly true at redshifts $1.90 < z_{\mathrm{LAE}} < 2.06$, since at these redshifts, [O II] $\lambda 3727$ would need to be blueshifted to be confused with Ly α. As mentioned, for $z_{\mathrm{LAE}} < 2.22$, the very small redshift of the [O II] emitters ($z_{[\mathrm{O\,II}]} < 0.05$) would likely allow their classification via their physical sizes and appearance on broad-band images. Thus, while studying the full redshift range of HETDEX is a perfectly valid approach, we would also like to see what would happen if we restricted measurements to the range over which [O II] galaxies are included in our simulations ($z_{[\mathrm{O\,II}]} > 0.05$). Removing the redshift range over which the LAE sample is pure means the contamination fraction for LAEs increases to 6.8 per cent for the high-contamination

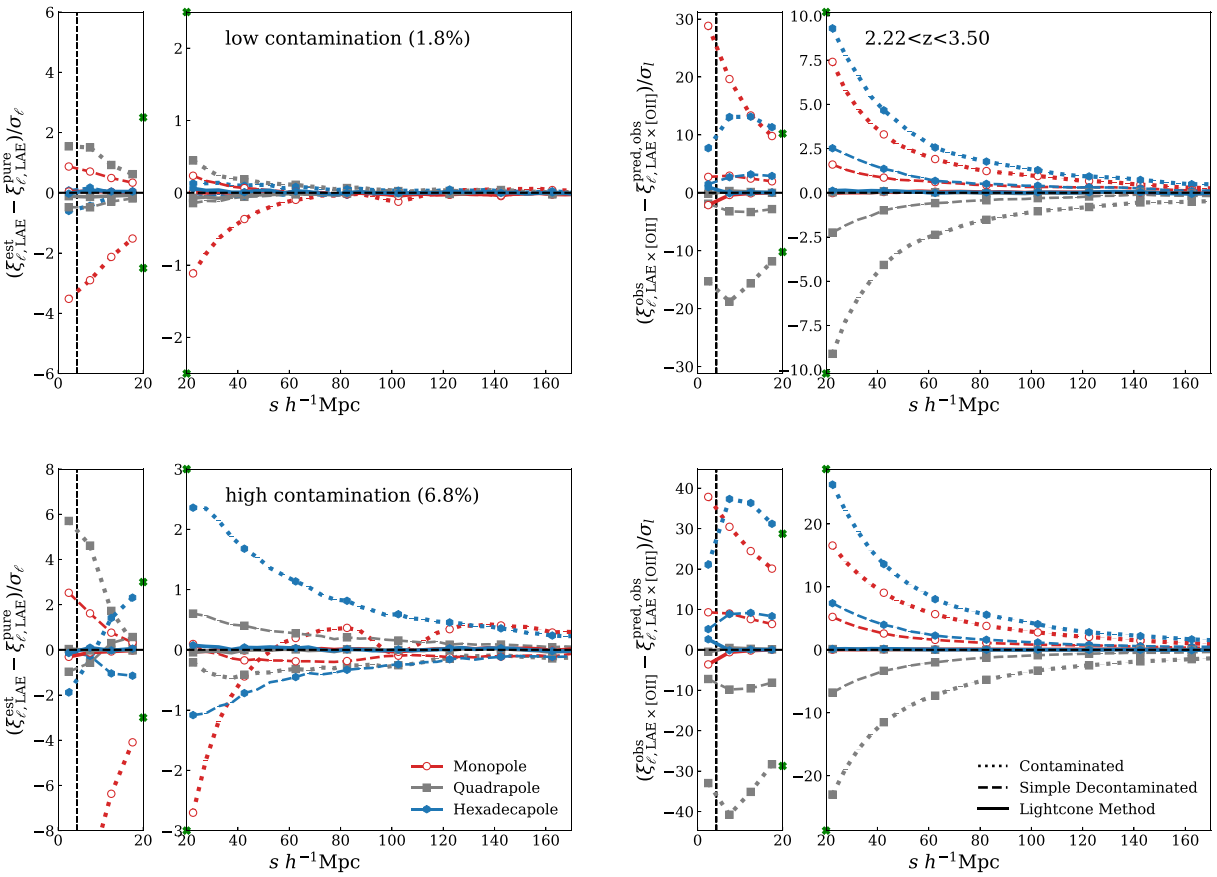

Figure 6. The same relations shown in Fig. 5, except now the redshift is restricted to the range over which [O II] emitters are simulated ($z_{[\mathrm{O\,II}]} > 0.05$). The results are very similar, though the residual spurious cross-correlation signal seen when ignoring the redshift dependence on the contamination is decreased.

sample (defined by $P_{\mathrm{LAE}} > 0.15$) and 1.8 per cent for the low-contamination case.

In Fig. 6, we show results for the redshift range $2.22 < z < 3.5$. As before the plots with autocorrelations show the difference between the (de)contaminated measurements and those from the corresponding pure catalogues with the same redshift range. In the new case, the contaminated cross-correlation function looks the same as for the full redshift range. However, the simple decontamination procedure, which ignores the redshift dependence of the purity within the redshift bin, works better than for the full z range. None the less, it can be seen from Fig. 6 that even when cutting out redshifts with the most dramatic changes in purity and contamination, not accounting for lightcone effects can still cause biases. In the lower contamination case, the simple decontamination leaves an $\sim 2\sigma$ biased cross-correlation monopole at separations $s > 20\,h^{-1}\,\mathrm{Mpc}$. This bias increases to $\sim 5\sigma$ for the higher contamination case. The autocorrelation also displays significant biases in the high-contamination case if redshift effects are ignored. The hexadecapole shows a systematic bias even at large scales of up to $\sim 1\sigma$, while the monopole shows an $\sim 2\sigma$ bias at the resolution limit of the simulation.

In contrast, as for the full redshift range analysis, the new lightcone-based decontamination method returns very close to the true autocorrelation function down to twice the cell size of the LAE box, and also accurately predicts the cross-correlation, with only tiny insignificant residuals, over the same range of scales.

6 FITTING THE CONTAMINATION

6.1 Fitting model and technique

The work so far has assumed we have perfect knowledge of the contamination. We now attempt to fit for the contamination by minimising the residual cross-correlation function, which as mentioned previously, should be zero for the case of perfect decontamination as the LAE and [O II] samples are in completely separate volumes. We do this using both the simple method and our lightcone-based approach to decontamination. We minimize the residual difference between the observed cross-correlation multipoles and the predicted cross-correlation multipoles evaluated using equations (19) and (27) for the simple and lightcone methods, respectively:

$$\chi^2 = \left(\boldsymbol{\xi}^{\mathrm{obs}}_{\mathrm{LAE}\times[\mathrm{O\,II}]} - \boldsymbol{\xi}^{\mathrm{pred,obs}}_{\mathrm{LAE}\times[\mathrm{O\,II}]}\right)^T \times \mathbf{C}^{-1} \left(\boldsymbol{\xi}^{\mathrm{obs}}_{\mathrm{LAE}\times[\mathrm{O\,II}]} - \boldsymbol{\xi}^{\mathrm{pred,obs}}_{\mathrm{LAE}\times[\mathrm{O\,II}]}\right), \quad (28)$$

where $\boldsymbol{\xi}_{\mathrm{LAE}\times[\mathrm{O\,II}]}$ is a vector of all the decontaminated multipole measurements, with superscripts indicating the observed ('obs') and predicted ('pred, obs') cross-correlation functions due to contamination, and $\mathbf{C}$ is the covariance matrix of 'observed' cross-correlation functions that we measure from the mock realizations. The data to which we fit our model are the mean of the 1000 measured cross-correlation functions, but we use the covariance matrix, $\mathbf{C}$, appropriate for a single realization. This means our results will have the reported uncertainty appropriate for a single HETDEX

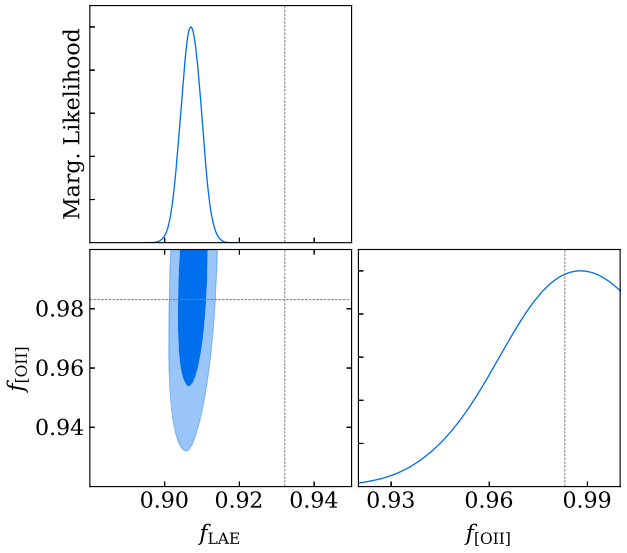

Figure 7. The MCMC 68 and 95 per cent contours for the simple decontamination method fits to the contaminated cross-correlation multipoles for the high contamination, $P_{LAE} > 0.15$ sample. The plots assume a single parameter purity model for LAE and [O II] galaxies. The dotted lines show the true purity measured directly from the mock catalogues. The normalized and marginalized 1D likelihoods are also shown.

realization, but they will be centred much closer to the best model description than statistically likely for real data.

In the lightcone method for the redshift dependence of the decontamination, we use a two-parameter model for each sample, where we fit the purity at the high and low edges of the redshift bin $f(z_{low})$ and $f(z_{high})$ for both LAEs and [O II] galaxies, corresponding to four parameters in total. We linearly interpolate between the two purity fractions to return purity values for the intermediate redshifts. This is motivated by considering the simplest model that could fit the redshift dependent contamination fractions shown in Fig. 3.

We need models of the LAE and [O II] clustering to make predictions for the cross-correlation for given contamination parameters. One approach could be to use the measured LAE and [O II] correlation functions, decontaminated using the model whose χ^2 is being evaluated. Since our aim is to present a proof of concept, and highlight the sensitivity of these statistics to the redshift dependent $f(z)$, we avoid this extra complication by using the true [O II] and LAE correlation functions, taken as the mean measurement of the 1000 pure mocks of the Fall field. For the [O II] correlation function in the simple method we use the mean correlation function measured from 1000 pure [O II] catalogues using LAE redshifts.

If one did estimate autocorrelation functions from the data, then their errors would have to be accounted for in the fitting routine. For the simple method one could use equation A16 of Awan & Gawiser (2020), which is an expression for the covariance of the decontaminated auto and cross-correlation functions that propagates the errors on the observed auto and cross- correlation functions. For the lightcone method, we argue we can avoid these issues in future work by including models of the LAE and [O II] correlation functions in the fitting (see more discussion in Section 7). As we only use the covariance of the cross-correlation function in our fits, it means the uncertainties in our parameter estimates correspond to the unrealistically optimistic scenario of perfect measurements or knowledge of the LAE and [O II] autocorrelation functions.

We fit the sample in the restricted redshift range $2.22 < z_{LAE} < 3.50$, since we assume all lower redshift [O II] emitters can be correctly classified by imaging. We note even in the restricted redshift range there is an area that in practice has 100 per cent LAE sample purity, where many other [O II] emission lines also lie in the HETDEX spectral range. Because of this, and the presence of other sharp features in the purity versus redshift relation, our simple straight line model for the contamination is not an ideal model of $f(z)$. Nevertheless, we shall see it does a reasonable job of describing the behaviour of contaminants.

To explore the posterior of the contamination model parameters, we use the software package COBAYA (Torrado & Lewis 2021), which uses the MCMC sampler of Lewis & Bridle (2002) and Lewis (2013). We use a flat prior between $0.6 < f < 1.0$ for the purity parameters, and compute the likelihood as $\mathcal{L} = \exp(-\chi^2/2)$. We ran eight chains for each fit and set the convergence criteria in COBAYA to require that a value of the Gelman–Rubin statistic (Gelman & Rubin 1992; Brooks & Gelman 1998), modified as described in Lewis (2013), of $R - 1 = 0.005$ is achieved (the COBAYA parameter RMINUS1_STOP). The smallest separation bin we use is $15.0 < s[h^{-1}\,\mathrm{Mpc}] < 20.0$. Even though we assume perfect knowledge of the LAE and [O II] autocorrelation functions, we avoid using smaller scales in the fit. This mimics a scenario where a model of the autocorrelation function might not work sufficiently well on small, more non-linear scales. The MCMC contour plots, confidence intervals, and best-fitting parameters were derived using the GETDIST Python package (Lewis 2019).

6.2 Results of the fit from the simple method

The simple method results of our MCMC fits to the cross-correlation function of the low-contamination LAE sample yield the sample purities $\tilde{f} = 0.975 \pm 0.003$ for LAEs and $\tilde{f}_{[O\,II]} = 0.96 \pm 0.03$ for [O II] emitters. The tildes specify quantities estimated from the MCMC fitting. We quote the best-fitting value and the range between the maximum and minimum value within 68 per cent of the highest likelihood weighed chain values. In Section 6.4 we explain why we do this rather than quoting the marginalized mean and standard deviation. These values can be compared to the total purity and contamination of the whole sample. The LAE purity has a slight bias with respect to the true purity of the sample ($f = 0.982$), while the [O II] purity value agrees with the true value of $f_{[O\,II]} = 0.956$ (these true values are exact to the given number of significant figures.) Note that since we fit to the mean of the mock measurements, the detection of this bias in LAE purity is more significant than its size.

The biases from the simple method of fitting for contamination are small in the low-contamination sample. In the high-contamination case (where 6.8 per cent of the LAE sample in the restricted redshift range are [O II] emitters), the bias on the LAE contamination is larger. We plot the results of the MCMC chains in Fig. 7. We can see from the figure that the LAE purity is much better constrained than the [O II] purity; the best-fitting LAE and [O II] purity values are $\tilde{f} = 0.907 \pm 0.004$ and $\tilde{f}_{[O\,II]} = 0.99^{+0.01}_{-0.03}$, respectively. For comparison, the true purity values are $f = 0.932$ and $f_{[O\,II]} = 0.983$. In this case, the LAE purity measurement is biased much lower than the truth, which is because the true purity values leave spurious cross-correlation signals (see Fig. 6), so a better χ^2 is given by biased parameters that return a smaller cross-correlation signal.

6.3 Results of the fit using the lightcone method

We now turn our attention to the new lightcone decontamination model, which uses two parameters to describe the redshift dependence of the sample purity. The results of our MCMC fitting of

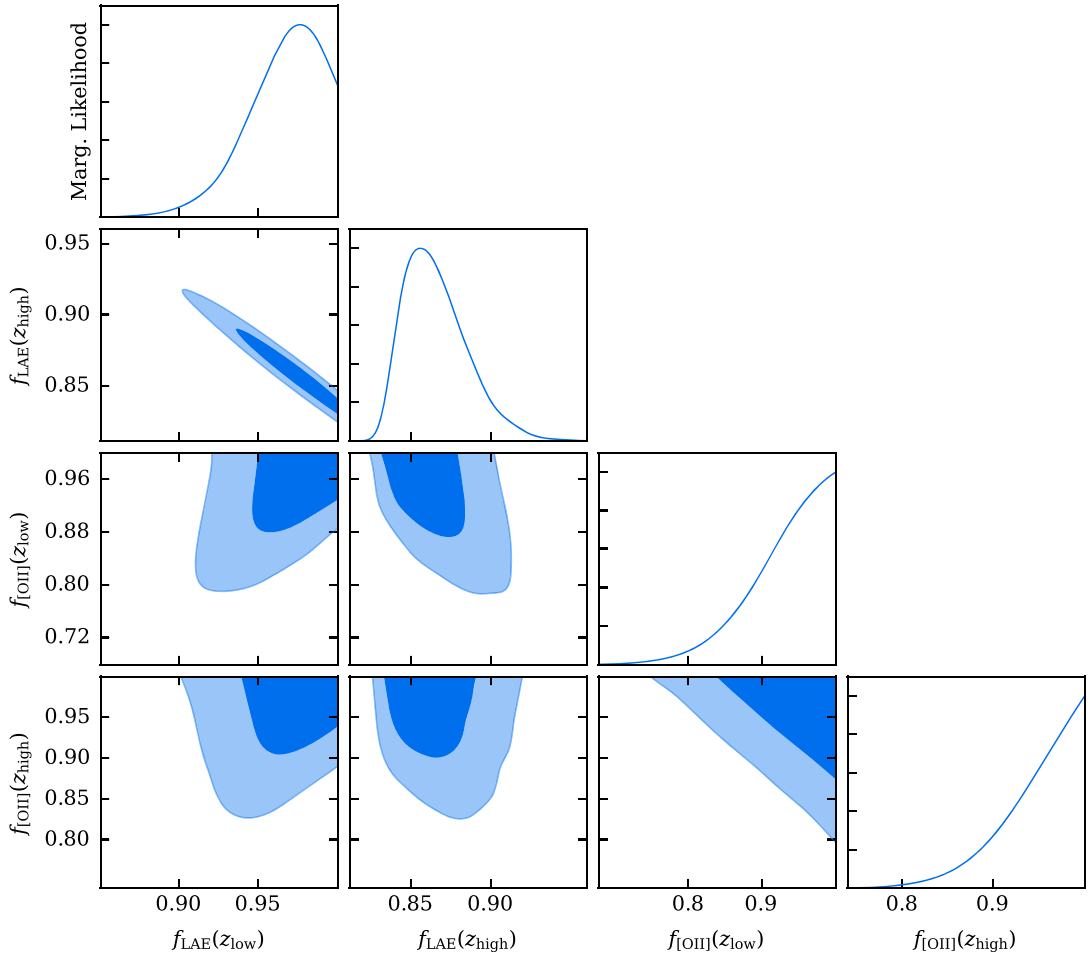

Figure 8. The 68 and 95 per cent MCMC contours derived by fitting the cross-correlation function of the mock catalogues using a two parameter model for the purity to the LAE and [O II] samples. These results are from our lightcone decontamination method. The data are for the high-contamination $P_{\rm LAE} > 0.15$ sample, which has an average contamination rate of 7 per cent. The normalized, marginalized 1D likelihoods are also shown. A redshift dependence of the contamination of the LAEs is clearly detected.

the cross-correlation function for the high-contamination sample are given in Fig. 8. We can see a degeneracy between the low and high redshift purity limits. Lower high-redshift completeness can be somewhat mitigated by higher low-redshift completeness. However, this degeneracy is not complete, and more contamination at higher redshifts is favoured. The [O II] emitters, on the other hand, are consistent with 100 per cent purity and there is no significant detection of any redshift dependence of the contamination.

The MCMC chains of the low-contamination sample also show the contamination of the LAEs is detected. In this case, however, no redshift dependence on purity is found. For the [O II] sample, 100 per cent purity is disfavoured but there is a degeneracy in the $f_{\rm [O\,II]}(z_{\rm low})$ and $f_{\rm [O\,II]}(z_{\rm high})$ parameters, which means the slope of the purity-redshift relation is not well constrained. Unlike in the simple method, there is no true value to compare the best-fitting parameters to (or to include in Fig. 8), as a straight line is not a perfect model of the true $f(z)$ or $f_{\rm [O\,II]}(z)$. Instead, in Section 6.4, we compare the straight-line fit from this method directly to the true purity versus redshift relations.

6.4 Constraints on the purity

We now wish to visualize our constraints in a plot of purity versus redshift. The flat priors on the purity at the ends of the redshift range result in non-flat priors on the derived values of purity at intermediate redshifts. For example, at the centre of the redshift range a contamination fraction in the middle of the prior range is favoured, as there are more allowed values of $f(z_{\rm low})$ and $f(z_{\rm high})$ that produce a purity crossing that point. In addition, we have the added complication that the expected purity is very close to the priors we use, and those priors cannot be expanded without including nonsensical values of purity (i.e. >1.0). Therefore, to minimize the effects of priors when plotting the purity constraints of Fig. 3, we do not plot the weighted mean and standard deviation of the chains. Instead we show the best-fitting parameters as a dotted line, and the maximum and minimum values of purity in 68 per cent of weighted chain values with the highest likelihoods as a shaded region. We compute these parameters using the likelihood functions in GETDIST (Lewis 2019).

We can see in Fig. 3 that for both of the $P_{\rm LAE}$ cuts, and for both the LAE and [O II] galaxy samples, the best-fitting parameters qualitatively reproduce the slope and amplitude that would be expected for a straight line fit to the more complex behaviour of the true purity. The true purity as a function of redshift is nearly always within the minimum and maximum range defined by the 68 per cent of chain positions with the highest likelihood. We do however see the true purity is slightly outside the region at the two redshift extremes. This is reasonable, as at those positions the true

purity deviates most from the straight line. It is possible the fits could be improved by adding additional parameters to the model to mimic the sharp drops in purity. However, we will see in Section 6.5 that this detailed modelling is unnecessary to decontaminate the clustering measurements for the scenarios considered here. It is also unclear whether the data are really constraining enough to warrant additional model parameters.

In general, we see that the LAE purity is better constrained than the [O II] purity. A possible cause of this is the choice of measuring the cross-correlation function using LAE redshifts for both samples. This mapping between observed wavelength and redshift means that the LAE contamination in the [O II] sample has a correlation function that is fixed with redshift. This allows us to remove it from the integral in equation (27). Since the LAE clustering term is independent of redshift, the sensitivity to the redshift dependence of purity in equation (27) comes from the $f_{[\mathrm{O\,II}]}(z)(1 - f(z))$ term. This term is most sensitive to changes in $f(z)$, since $f_{[\mathrm{O\,II}]}(z)$ is always close to one but $(1 - f(z))$ is always very small. One way to better constrain the [O II] purity versus redshift could be to measure the cross-correlation using [O II] redshifts for all sources. Then the projected LAE clustering would evolve over the redshift range of [O II], making the cross-correlation more sensitive to $f_{[\mathrm{O\,II}]}(z)$. Note this approach is not guaranteed to give a strong constraint on the [O II] purity, as the LAE sample has an intrinsically weaker correlation function, which may get even weaker when projected to lower redshifts. Since we are mainly interested in the contamination of the LAE correlation function, and $f(z)$ is the only purity term appearing in equation (17), we leave further study of this to future work.

6.5 Using the fitted contamination

In Section 5, we assume we have perfect knowledge of the purity as a function of redshift. Here we use the purity we have fit from the cross-correlation function, to see if it can be used to give unbiased, decontaminated measurements of the autocorrelation functions. We decontaminate the measurements from the two fields separately using the estimated parameters from the combined field, and then combine them after the decontamination (following, e.g. White et al. 2011). To test if combining before or after decontamination makes a difference in our mocks, we tried combining the two fields before decontaminating for one of the scenarios, specifically the high-contamination sample using the simple decontamination method. We found results that agree closely, confirming the order of combining and decontaminating is not important for our mocks. In the future, further tests could be carried out to find if this is also the case with the real HETDEX data.

In Fig. 9, we show the results of using the best-fitting contamination parameters. We can see that these parameters return much smaller residual cross-correlation signals than the true parameters for the simple approach that ignores redshift effects. This is because it is able to fit the extra cross-correlation signal with artificially low purity values. However, the LAE autocorrelation function shows the danger in this, as the incorrect inferred purity values results in even more biased results than when using the true values. The monopole from the high-contamination sample can have a $\sim 2\sigma$ bias even at larger scales ($s > 20\,h^{-1}$ Mpc) and the bias gets even worse at small scales. The low-contamination sample monopole shows a $\sim 1\sigma$ bias at large scales, increasing to $\sim 2\sigma$ at small scales.

In contrast, our new lightcone method of accounting for the redshift dependence of contamination works well for both the auto and cross-correlation functions, yielding autocorrelation measurements with only an $\sim 0.3\sigma$ bias at large scales ($s > 20\,h^{-1}$ Mpc) and biases smaller than the statistical error down to twice the cell size of the simulation box. The results from the fitted contamination are slightly worse than the results from using the true purity versus redshift, but this is to be expected, given the limitations of the model parametrization and the added uncertainties from the fit. These results show this method could potentially be used to both fit for contamination and correct clustering measurements.

It is important to note that in Fig. 9 we have the benefit of a best-fitting purity measured from our large number of mock catalogues. In the real data, there will be some statistical error associated with the best-fitting purity parameters that would have to be propagated into the final results.

It should also be noted that an alternative method of deriving the $f(z)$ to the one presented here is to use the simple decontamination method in narrow redshift bins where the contamination and projection parameters are roughly constant. This could have benefits, such as avoiding potentially losing information by not integrating over redshift, and not needing a model of the LAE or [O II] correlation function. On the other hand, using narrow redshift bins would make the individual measurements noisier. We leave a comparative study of these approaches for later work. We also highlight regardless of how the $f(z)$ is derived, the lightcone formalism allows one to optimize the size of redshift bins for the cosmology measurements without having to be restricted by the requirement of avoiding projected interloper clustering evolution.

7 CONCLUSIONS

We generated 1000 mock catalogues of the HETDEX survey, which include clustering, redshift-space distortions, redshift-dependent noise, and a realistic selection function. We used a reformulated version of the probabilistic classifier of Leung et al. (2017) to generate catalogues of LAEs with realistic, redshift-dependent [O II] galaxy contamination, and considered two scenarios that bracket the expectations for HETDEX: low (1–2 per cent) and high contamination (5–7 per cent). These catalogues were used to explore the impact of the redshift dependence of the contamination fraction and the correlation function of the contaminants on the observed correlation functions.

The mock catalogues show that existing methods of decontamination such as Awan & Gawiser (2020), and other methods that do not account for redshift evolution of the interlopers within a redshift bin, should not be directly applied to clustering measurements from a survey such as HETDEX, unless the analysis is restricted to using redshift bins that are narrow enough for the evolution effects to not be important – which is unlikely to be ideal for the HETDEX cosmological analysis. This is because in HETDEX both the [O II] galaxy contamination fraction and the projected [O II] clustering vary with redshift. Although in the low-contamination cases we consider, such methods may be effective enough, when the contamination is larger (with misclassified fractions of 5–7 per cent) biases appear. In the autocorrelation function, these biases can be larger than the statistical error of HETDEX. Moreover, the biases in the cross-correlation signal are even stronger, with a spurious signal of more than 5σ of the expected statistical noise being left after subtracting the expected cross-correlation from contamination. A biased, decontaminated cross-correlation function is not a problem for HETDEX in itself. However, we have also shown the inability of the simple decontamination method to correctly account for the spurious cross-correlation signal results in biases in the inferred contamination

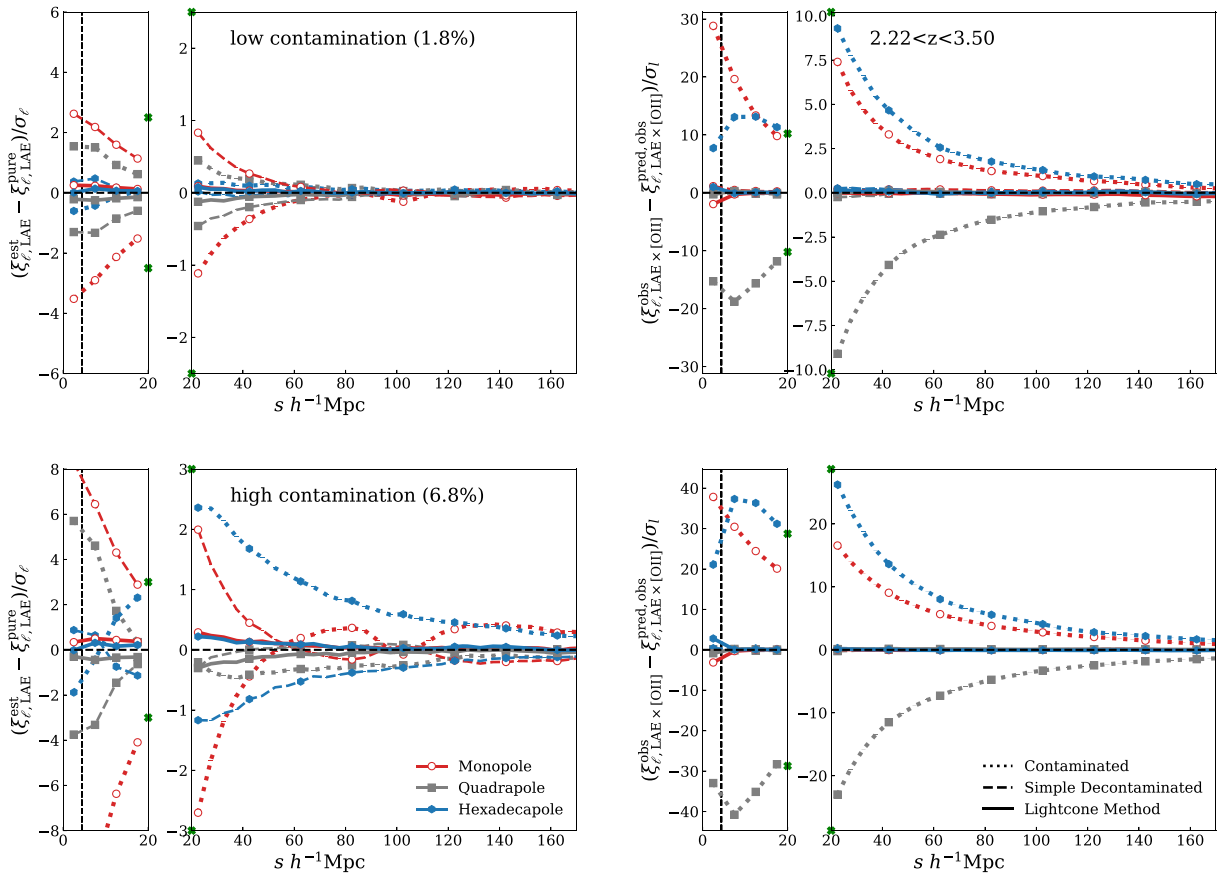

Figure 9. This same plot as in Fig. 6, except now instead of using the true purity as a function of redshift, the purity values are determined by fitting the cross-correlation functions. In this case all decontamination methods are forced towards returning a zero residual cross-correlation signal in order to minimize the χ^2. However, if we ignore the redshift dependence of the contamination, the best-fitting contamination value is too high; this error propagates directly into additional biases in the decontaminated LAE autocorrelation function. The new biases are most noticeable in the higher contamination results shown on the left-hand panel of the bottom row. The new lightcone approach however works well for both the auto and cross-correlation functions.

fractions from fits to the observed, raw cross-correlation functions. If these incorrect contamination fractions are used to decontaminate the autocorrelation function, additional biases will be propagated into the measurements used to fit cosmological parameters.

We present a method to account for the redshift effects that can be applied when there is no true cross-correlation between the pure samples of the target galaxies and the contaminants. Our method combines the literature approach to decontamination with the models of correlation functions integrated along a lightcone given in Yamamoto & Suto (1999) and Suto et al. (2000). This methodology is needed in scenarios like HETDEX, where the correlation function of the [O II] interlopers evolves rapidly due to projection effects. Accounting for the lightcone effects gives a much better model of the cross-correlation, and it also produces decontaminated autocorrelation functions that agree with the pure measurements to an accuracy much smaller than the statistical noise down to 20 h^{-1} Mpc. Although we formulate this work for the correlation function, our findings should also apply to the power spectrum.

The work on this topic is not complete. The method we have developed is an improvement over existing methods, but we still have to assume that the true clustering of LAEs and [O II] galaxies does not evolve with redshift. Allowing for evolving LAE and [O II] correlation functions is possible in the framework we present however, and such an evolution could be constrained with autocorrelation function measurements from the data. We also mention once more that we always compute the projected clustering using distortion parameters assuming a true cosmology. As advocated by Addison et al. (2019), future work could consider the impact of this limitation.

Although our decontamination method itself relies only on interpolating over the observed correlation functions and the assumption of a fiducial cosmology for computing the distortion parameters, our method of fitting the contamination assumes perfect knowledge of the LAE and [O II] autocorrelation functions. One way to make our experiments with fitting the contamination applicable to real data would be to develop an approach that simultaneously fits models of the cosmology, bias, and the contamination parameters to both the cross- and autocorrelation functions. The distortion parameters could also be modified to be consistent with the different cosmologies during the fit, solving a further issue.

Simultaneously fitting contamination, galaxy bias, and cosmology is advocated in e.g. Addison et al. (2019) and Grasshorn Gebhardt et al. (2019). While their approaches assumed single contamination values for the whole sample, we would suggest including our new redshift dependent modelling of contamination. This is especially true if the contamination fraction of HETDEX is closer to 5 per cent than 1 per cent, and if there are significant differences in the sample purity at different redshifts. Before applying the method to real data, a modelling pipeline that includes redshift-dependent contamination should be tested on more realistic simulations than our lognormal mocks. These simulations should include possible evolution in the true galaxy correlation function and a better modelling of the clustering and redshift space distortions on non-linear scales.

To summarize, this paper highlights the importance of the redshift dependence of the contamination, presents a method to model these effects, and shows that such effects should be considered when decontaminating clustering measurements from surveys with redshift-dependent contamination within the adopted redshift bins. These effects are particularly important when using the cross-correlation function to constrain contamination.

ACKNOWLEDGEMENTS

The authors acknowledge the feedback from the internal HETDEX referees and the anonymous journal referee. We acknowledge useful discussions with Humna Awan, Jiamin Hou, Martha Lippich, Andrea Pezzotta, Agne Semenaite, Martín Crocce, Román Scoccimarro, and the HETDEX cosmology science working group. Henry S. Grasshorn Gebhardt is a National Aeronautics and Space Administration (NASA) Postdoctoral Program Fellow. KG acknowledges support from the National Science Foundation (NSF) via grant NSF-2008793. EG was supported by the Department of Energy via grant DE-SC0010008. EK's work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2094 – 390783311. DJ was supported at Pennsylvania State University by the NASA ATP program (80NSSC18K1103). We acknowledge the use of the PYTHON libraries MATPLOTLIB (Hunter 2007), ASTROPY (Astropy Collaboration et al. 2013, 2018), NUMPY (Harris et al. 2020), and SCIPY (Virtanen et al. 2020). This research also used TOPCAT (Taylor 2005), STILTS (Taylor 2006); and the GNU Scientific Library (GSL); URL: https://www.gnu.org/software/gsl/

HETDEX is led by the University of Texas at Austin McDonald Observatory and Department of Astronomy with participation from the Ludwig-Maximilians-Universität München, Max-Planck-Institut für Extraterrestrische Physik (MPE), Leibniz-Institut für Astrophysik Potsdam (AIP), Texas A&M University, Pennsylvania State University, Institut für Astrophysik Göttingen, The University of Oxford, Max-Planck-Institut für Astrophysik (MPA), The University of Tokyo, and Missouri University of Science and Technology. In addition to Institutional support, HETDEX is funded by the National Science Foundation (grant AST-0926815), the State of Texas, the US Air Force (AFRL FA9451-04-2- 0355), and generous support from private individuals and foundations.

The observations were obtained with the Hobby-Eberly Telescope (HET), which is a joint project of the University of Texas at Austin, the Pennsylvania State University, Ludwig-Maximilians-Universität München, and Georg-August-Universität Göttingen. The HET is named in honour of its principal benefactors, William P. Hobby and Robert E. Eberly.

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high performance computing, visualization, and storage resources that have contributed to the research results reported within this paper; URL: http://www.tacc.utexas.edu

This research made use of NASA's Astrophysics Data System Bibliographic Services.

DATA AVAILABILITY

The HETDEX data are currently proprietary, but public releases are planned for the future. The authors will respond to reasonable requests for access to the simulation data used in this paper, so long as no unreleased proprietary data is involved.

REFERENCES

Abbott T. M. C. et al., 2018, ApJS, 239, 18
Acquaviva V., Gawiser E., Guaita L., 2011, ApJ, 737, 47
Addison G. E., Bennett C. L., Jeong D., Komatsu E., Weiland J. L., 2019, ApJ, 879, 15
Agrawal A., Makiya R., Chiang C.-T., Jeong D., Saito S., Komatsu E., 2017, J. Cosmol. Astropart. Phys., 2017, 003
Alcock C., Paczynski B., 1979, Nature, 281, 358(AP)
Anders P., Fritze-v. Alvensleben U., 2003, A&A, 401, 1063
Astropy Collaboration et al., 2013, A&A, 558, A33
Astropy Collaboration et al., 2018, AJ, 156, 123
Awan H., Gawiser E., 2020, ApJ, 890, 78
Ballinger W. E., Peacock J. A., Heavens A. F., 1996, MNRAS, 282, 877
Behrens C., Byrohl C., Saito S., Niemeyer J. C., 2018, A&A, 614, A31
Blake C., Pope A., Scott D., Mobasher B., 2006, MNRAS, 368, 732
Brooks S. P., Gelman A., 1998, J. Comput. Graph. Stat., 7, 434–455
Byrohl C., Saito S., Behrens C., 2019, MNRAS, 489, 3472
Byrohl C. et al. 2021, MNRAS, 506, 5129
Catelan P., Lucchin F., Matarrese S., Porciani C., 1998, MNRAS, 297, 692
Catelan P., Porciani C., Kamionkowski M., 2000, MNRAS, 318, L39
Chan K. C., Scoccimarro R., Sheth R. K., 2012, Phys. Rev. D, 85, 083509
Cheng Y.-T., Chang T.-C., Bock J., Bradford C. M., Cooray A., 2016, ApJ, 832, 165
Cheng Y.-T., Chang T.-C., Bock J. J., 2020, ApJ, 901, 142
Chiang C.-T. et al., 2013, J. Cosmol. Astropart. Phys., 12, 030
Ciardullo R. et al., 2013, ApJ, 769, 83
Crocce M., Scoccimarro R., 2006, Phys. Rev. D, 73, 063519
Dawson K. S. et al., 2013, AJ, 145, 10
Desjacques V., Jeong D., Schmidt F., 2018, Phys. Rep., 733, 1
Doi M. et al., 2010, AJ, 139, 1628
Eggemeier A., Scoccimarro R., Crocce M., Pezzotta A., Sánchez A. G., 2020, Phys. Rev. D, 102, 103530
Fitzpatrick E. L., 1999, PASP, 111, 63
Flaugher B. et al., 2015, AJ, 150, 150
Fry J. N., 1996, ApJ, 461, L65
Gelman A., Rubin D. B., 1992, Stat. Sci., 7, 457
Gong Y., Silva M., Cooray A., Santos M. G., 2014, ApJ, 785, 72
Gong Y., Chen X., Cooray A., 2020, ApJ, 894, 152
Grasshorn Gebhardt H. S. et al., 2019, ApJ, 876, 32
Green G., 2018, J. Open Source Softw., 3, 695
Gronwall C., Ciardullo R., Matkovic A., Feldmeier J. J., Hay J., MUSYC Collaboration, 2014, AAS Meeting Abstracts #223, p. 246.39
Gurung-López S., Orsi Á. A., Bonoli S., Baugh C. M., Lacey C. G., 2019, MNRAS, 486, 1882
Gurung-López S., Orsi Á. A., Bonoli S., Padilla N., Lacey C. G., Baugh C. M., 2020, MNRAS, 491, 3266
Hamilton A. J. S., Tegmark M., 2004, MNRAS, 349, 115
Harris C. R. et al., 2020, Nature, 585, 357
Hill G. J. et al., 2008, in Kodama T., Yamada T., Aoki K., eds, ASP Conf. Ser. Vol. 399, Panoramic Views of Galaxy Formation and Evolution. Astron. Soc. Pac., San Francisco, p. 115,
Hill G. J. et al., 2018, in Evans C. J., Simard L., Takami H., eds, SPIE Conf. Ser. Vol. 10702, Ground-based and Airborne Instrumentation for Astronomy VII. SPIE, Bellingham, p. 107021K
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Kaiser N., 1987, MNRAS, 227, 1
Kawanomoto S. et al., 2018, PASJ, 70, 66
Khostovan A. A. et al., 2019, MNRAS, 489, 555
Landy S. D., Szalay A. S., 1993, ApJ, 412, 64
Lazeyras T., Wagner C., Baldauf T., Schmidt F., 2016, J. Cosmol. Astropart. Phys., 2016, 018
Leung A. S. et al., 2017, ApJ, 843, 130
Lewis A., 2013, Phys. Rev. D, 87, 103529
Lewis A., 2019, preprint (arXiv:1910.13970)
Lewis A., Bridle S., 2002, Phys. Rev. D, 66, 103511
Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473
Lidz A., Taylor J., 2016, ApJ, 825, 143

Madau P., 1995, ApJ, 441, 18
Massara E., Ho S., Hirata C. M., DeRose J., Wechsler R. H., Fang X., 2020, preprint (arXiv:2010.00047)
Munari U., Sordo R., Castelli F., Zwitter T., 2005, A&A, 442, 1127
Oke J. B., Gunn J. E., 1983, ApJ, 266, 713
Papovich C. et al., 2016, ApJS, 224, 28
Peebles P. J. E., 1980, The Large-scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Planck Collaboration et al., 2020, A&A, 641, A6
Pullen A. R., Hirata C. M., Doré O., Raccanelli A., 2016, PASJ, 68, 12
Quadri R. et al., 2007, ApJ, 654, 138
Sánchez A. G., 2020, Phys. Rev. D, 102, 123511
Sánchez A. G. et al., 2017, MNRAS, 464, 1640
Schlafly E. F. et al., 2010, ApJ, 725, 1175
Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Suto Y., Magira H., Yamamoto K., 2000, PASJ, 52, 249
Swanson M. E. C., Tegmark M., Hamilton A. J. S., Hill J. C., 2008, MNRAS, 387, 1391
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, ASP Conf. Ser. Vol. 347, Astronomical Data Analysis Software and Systems XIV. Astron. Soc. Pac., San Francisco, p. 29
Taylor M. B., 2006, in Gabriel C., Arviset C., Ponz D., Enrique S., eds, ASP Conf. Ser. Vol. 351, Astronomical Data Analysis Software and Systems XV. Astron. Soc. Pac., San Francisco, p. 666
Torrado J., Lewis A., 2021, J. Cosmol. Astropart. Phys., 2021, 057
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Visbal E., Loeb A., 2010, J. Cosmol. Astropart. Phys., 2010, 016
Wagner C., Schmidt F., Chiang C. T., Komatsu E., 2015, MNRAS, 448, L11
White M. et al., 2011, ApJ, 728, 126
Wold I. G. B. et al., 2019, ApJS, 240, 5
Yamamoto K., Suto Y., 1999, ApJ, 517, 1

APPENDIX A: TESTS OF THE PROJECTION

In this section, we further test our modelling of the lightcone projection and our use of the distortion parameters $c_\perp$ and $c_\parallel$. To do this, we measure the projected clustering of pure samples of [O II] emitters in 1000 of our Fall field mock catalogues. This means we measure the multipoles of a pure [O II] catalogue assuming Ly α redshifts for everything. To assess how well equation (21) models the redshift dependent distortion of the [O II] density field, we compute a prediction of the projected correlation function for the redshift interval $2.22 < z < 2.5$ via

$$\xi^{\rm proj}(s,\mu) = \mathcal{F}\left[n^{\rm pure}_{\rm [O\,II]}(z), \xi^{\rm proj}_{\rm [O\,II]}(s,\mu,z)\right], \tag{A1}$$

where $n^{\rm pure}_{\rm [O\,II]}(z)$ is the number density as a function of LAE redshift of the projected, pure [O II] sample. We then compare the multipoles of the result to the multipoles measured from the mock catalogue $\xi^{\rm mock}_\ell(s)$.

The difference between the mock and predicted multipoles of the projected [O II] galaxy correlation function is given as dashed lines in Fig. A1; the measurements are divided by the statistical error expected for one HETDEX Fall-field realization. We can see only very small ($\sim 0.2\sigma$) differences between the statistical errors down to $s = 20\,h^{-1}$ Mpc, which gives further confirmation of our use of the Yamamoto & Suto (1999) and Suto et al. (2000) approach to account for the redshift dependence of the projection parameters. On scales smaller than this, the differences increase, becoming approximately the same size as the statistical error around $s = 10\,h^{-1}$ Mpc. It is unclear why the modeling of the projection does not work perfectly down to the smallest scales, but for the purposes of decontaminating the LAE signal, the projected [O II] clustering is down weighted by the square of the LAE sample contamination. Thus, this small bias should not affect our results on LAE clustering decontamination. In Section 6, where we fit the cross–correlation, we restrict ourselves to larger separations that will also mitigate any possible effect.

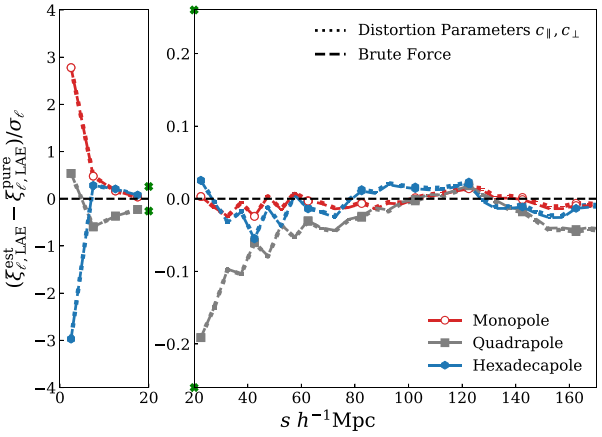

Figure A1. The difference between the mean multipoles measured from 1000 Fall field mock catalogues of [O II] emitters, analysed using redshifts assuming the [O II] emitters are actually LAEs, and our prediction of the projected [O II] galaxy clustering. Dashed lines use our standard approach with distortion parameters and an integral along the redshift range. The dotted lines replace the distortion parameters with the brute force approach detailed in the text. The distortion parameters and the brute force approach give nearly indistinguishable results. The plot is split into two panels to allow a larger dynamic range, in the right-hand panel only every forth data point is marked with a symbol for visual clarity.

In this paper, we use a model of $\xi^{\rm proj}_{\rm [O\,II]}(s,\mu,z)$ that relies on distortion parameters, i.e. equation (13). These distortion parameters are, however, an approximate model of the effects of the true projection as they do not account for the different redshifts of the two galaxies in the pair. We therefore also compute a mapping between the s and μ coordinates using a brute force method. In 800 uniform bins of LAE redshift within $2.06 < z < 3.5$ (i.e. the range where a Ly α emitter has a wavelength greater than 3727 Å), we generate pairs of galaxies in Cartesian coordinates with given values for the true separation s, μ and compute an observed Right Ascension and Declination to a virtual observer. The pairs we generate are always in a plane and use a fixed LOS. We then recompute the Cartesian coordinates from our 'observed' coordinates but assume [O II] redshifts, and recompute the LOS to the galaxy pair. By measuring the r', μ' of this projected pair we generate a look-up table between true and projected coordinates in 400 bins of r and μ. We interpolate over this 3D look-up table to provide an alternative coordinate mapping for our model of the projected [O II] clustering.

The dotted lines in Fig. A1 show the difference of the mock multipoles and the projected multipoles using the brute force look-up table for the mapping between LAE and [O II] coordinates. There is extremely close agreement between the predictions using the distortion parameters and the brute force look-up table. This confirms that the distortion parameters $c_\parallel$, $c_\perp$, which are advocated by several other authors (e.g. Visbal & Loeb 2010; Gong et al. 2014; Lidz & Taylor 2016; Pullen et al. 2016; Leung et al. 2017; Grasshorn Gebhardt et al. 2019), are an excellent model of the true distortion for the HETDEX LAE/[O II] confusion scenario.

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