

# Correlations in the three-dimensional Lyman-alpha forest contaminated by high column density absorbers

Keir K. Rogers,<sup>1,2★</sup> Simeon Bird,<sup>3†</sup> Hiranya V. Peiris,<sup>1,2</sup> Andrew Pontzen,<sup>2</sup>  
Andreu Font-Ribera<sup>2</sup> and Boris Leistedt<sup>4†</sup>

<sup>1</sup>*Oskar Klein Centre for Cosmoparticle Physics, Stockholm University, AlbaNova, Stockholm SE-106 91, Sweden*

<sup>2</sup>*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*

<sup>3</sup>*Department of Physics and Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA*

<sup>4</sup>*Department of Physics, New York University, 726 Broadway, New York, NY 10003, USA*

Accepted 2018 March 2. Received 2018 March 2; in original form 2017 November 30

## ABSTRACT

Correlations measured in three dimensions in the Lyman-alpha forest are contaminated by the presence of the damping wings of high column density (HCD) absorbing systems of neutral hydrogen (H I; having column densities  $N(\text{H I}) > 1.6 \times 10^{17}$  atoms  $\text{cm}^{-2}$ ), which extend significantly beyond the redshift-space location of the absorber. We measure this effect as a function of the column density of the HCD absorbers and redshift by measuring three-dimensional (3D) flux power spectra in cosmological hydrodynamical simulations from the Illustris project. Survey pipelines exclude regions containing the largest damping wings. We find that, even after this procedure, there is a scale-dependent correction to the 3D Lyman-alpha forest flux power spectrum from residual contamination. We model this residual using a simple physical model of the HCD absorbers as linearly biased tracers of the matter density distribution, convolved with their Voigt profiles and integrated over the column density distribution function. We recommend the use of this model over existing models used in data analysis, which approximate the damping wings as top-hats and so miss shape information in the extended wings. The simple ‘linear Voigt model’ is statistically consistent with our simulation results for a mock residual contamination up to small scales ( $|\mathbf{k}| < 1 h \text{ Mpc}^{-1}$ ). It does not account for the effect of the highest column density absorbers on the smallest scales (e.g.  $|\mathbf{k}| > 0.4 h \text{ Mpc}^{-1}$  for small damped Lyman-alpha absorbers; HCD absorbers with  $N(\text{H I}) \sim 10^{21}$  atoms  $\text{cm}^{-2}$ ). However, these systems are in any case preferentially removed from survey data. Our model is appropriate for an accurate analysis of the baryon acoustic oscillations feature. It is additionally essential for reconstructing the full shape of the 3D flux power spectrum.

**Key words:** quasars: absorption lines – large-scale structure of universe – cosmology: theory.

## 1 INTRODUCTION

Absorption lines of the Lyman-alpha forest can be mapped in three dimensions (i.e. line-of-sight direction along the lengths of quasar spectra and transverse direction in the angular positions of the spectra on the sky) to trace the fluctuations in the cosmological density field. Correlations in the Lyman-alpha forest are a powerful probe of high redshifts ( $z > 2$ ), before dark energy came to dominate the evolution of the Universe. In particular, measurement of the three-dimensional (3D) correlations on large scales (separations

$r \sim 100 \text{ Mpc } h^{-1}$ ) in the Lyman-alpha forest allows a measurement of the baryon acoustic oscillations (BAOs) in the distribution of matter at  $z \sim 2.3$  (Slosar et al. 2011; Busca et al. 2013; Kirkby et al. 2013; Slosar et al. 2013; Delubac et al. 2015; Bautista et al. 2017). 3D correlations between the Lyman-alpha forest and the distribution of quasars have also been measured, including the detection of BAOs (Font-Ribera et al. 2013, 2014a; du Mas des Bourboux et al. 2017). This has been achieved, thanks to the large number of quasar spectra from the Baryon Oscillation Spectroscopic Survey (BOSS; Eisenstein et al. 2011; Dawson et al. 2013, 157 783 were suitable for analysis in Data Release 12; DR12) and the large sky area they cover (the footprint in DR12 covers approximately one quarter of the sky). Consequently, Lyman-alpha forest analyses are no longer restricted to measurements of the 1D flux power spectrum

\* E-mail: keir.rogers@fysik.su.se

† NASA Einstein Fellow.

(along the line of sight only), which probes smaller scale clustering ( $k_{\parallel} > 0.1 h \text{Mpc}^{-1}$ ) and constrains cosmological models that suppress small-scale power, e.g. those containing massive neutrinos or warm dark matter (WDM; Seljak et al. 2005; Palanque-Delabrouille et al. 2015; Armengaud et al. 2017; Iršič et al. 2017a,b; Yeche et al. 2017).

Current measurements of the 3D correlations in the Lyman-alpha forest reconstruct the correlation function, where the BAO feature is most distinguishable. However, ongoing analyses in the extended Baryon Oscillation Sky Survey (eBOSS; Dawson et al. 2016) and future surveys like the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016a,b) will also measure its Fourier-space counterpart, the 3D flux power spectrum (Font-Ribera, McDonald & Slosar 2018). A measurement of the 3D Lyman-alpha forest power spectrum will probe the full shape on a wide range of scales ( $0.01 < k < 1 h \text{Mpc}^{-1}$ ). On large scales ( $k < 0.1 h \text{Mpc}^{-1}$ ), the 3D flux power spectrum can be used to determine the cosmological geometry through the Alcock-Paczynski test (Alcock & Paczynski 1979; Hui, Stebbins & Burlles 1999; McDonald & Miralda-Escudé 1999; McDonald 2003). The 3D forest power spectrum on large scales can also be used to study fluctuations in the ultraviolet ionizing background (Pontzen 2014; Pontzen et al. 2014).

On smaller scales ( $0.1 < k < 1 h \text{Mpc}^{-1}$ ), the 3D flux power spectrum adds complementary information to that from the 1D flux power spectrum. Font-Ribera et al. (2018) show that the 3D flux power spectrum for the BOSS survey is more constraining than the 1D counterpart up to a maximum  $k = 1 h \text{Mpc}^{-1}$ . Only for  $k > 1 h \text{Mpc}^{-1}$  does the 1D power spectrum contain essentially all information. For future surveys such as DESI, where there will be a higher density of lines of sight, one may anticipate 3D information to even higher  $k$ , underscoring the importance of working with the 3D spectrum wherever possible. This will provide more power to constrain cosmological models with additional components (e.g. massive neutrinos, WDM, or fuzzy dark matter), or modifications to a simple power-law primordial power spectrum (e.g. running of the primordial spectral index). In addition to providing greater statistical power, the 3D flux power spectrum is sensitive to different systematics than the 1D flux power spectrum (e.g. in correlations with metal absorption lines). (See e.g. Font-Ribera et al. 2014b for forecasts of the constraining power of the 3D flux power spectrum with DESI.)

As with the 1D Lyman-alpha forest flux power spectrum (McDonald et al. 2005; Rogers et al. 2018), 3D correlations in the Lyman-alpha forest are biased by the presence in quasar spectra of high column density (HCD) absorbers and their associated broadened absorption lines (McQuinn & White 2011; Slosar et al. 2011; Font-Ribera & Miralda-Escudé 2012). HCD absorbers are defined as regions of neutral hydrogen (H I) gas with a column density  $N(\text{H I})$  exceeding  $1.6 \times 10^{17} \text{ atoms cm}^{-2}$ , and are usually identified with the gas in or around galaxies. They form at the peaks of the underlying density distribution and so cluster more strongly than the Lyman-alpha forest (Font-Ribera et al. 2012; Pérez-Ràfols et al. 2018). The absorption lines of the highest column density systems are broadened, with large damping wings causing absorption in the spectrum away from the physical location of the absorber. These wings have a characteristic Voigt profile, a convolution of a Gaussian profile (caused by Doppler broadening), and a Lorentzian profile (caused by natural or collisional broadening). They are traditionally sub-classified as either damped Lyman-alpha absorbers (DLAs;  $N(\text{H I}) > 2 \times 10^{20} \text{ atoms cm}^{-2}$ ) or Lyman-limit systems ( $1.6 \times 10^{17} \text{ atoms cm}^{-2} < N(\text{H I}) < 2 \times 10^{20} \text{ atoms cm}^{-2}$ ), according to the width of their damping wings (Wolfe et al. 1986). However,

as noted in, for example, McDonald et al. (2005), Font-Ribera & Miralda-Escudé (2012), and Rogers et al. (2018), systems with  $N(\text{H I})$  exceeding  $1 \times 10^{19} \text{ atoms cm}^{-2}$  have significant wings, which we classify as sub-DLAs. In Lyman-alpha forest analyses, it is usual to attempt to ‘clip’ out HCD absorbers by identifying their damping wings in spectra, masking the central absorption region and then correcting the wings (e.g. see Lee et al. 2013 for details of the process for BOSS DR9 spectra). None the less, there is always a residual contamination of HCD absorbers, since the smallest damping wings are hard to identify amongst instrumental noise and indeed the superposed Lyman-alpha forest itself. Estimates of the upper limit in column density for this residual contamination range from  $10^{20}$  to  $10^{21} \text{ atoms cm}^{-2}$  (e.g. Bautista et al. 2017). It is therefore necessary to model the effect of this residual contamination to allow for robust cosmological inference from the Lyman-alpha forest (Bautista et al. 2017 and du Mas des Bourboux et al. (2017) were the first to model this component in a 3D correlation analysis).

There is a small literature on modelling the effect of (residual) HCD absorbers on correlations in the 3D Lyman-alpha forest. In appendix B of McQuinn & White (2011), a linear model for the 3D flux power spectrum of HCD absorbers convolved with their (Voigt) absorption profiles is considered, allowing for their auto-correlation and cross-correlation with the Lyman-alpha forest. They show that this ‘linear Voigt model’ predicts that the cross-correlation is the dominant component of the HCD absorbers’ correction to the 3D Lyman-alpha forest power spectrum. Font-Ribera & Miralda-Escudé (2012) measure the effect of HCD absorbers on the 3D Lyman-alpha forest correlation function using mock (quasar) spectra (details of their generation are given in Font-Ribera, McDonald & Miralda-Escudé 2012). They find the cross-correlation of HCD absorbers and the Lyman-alpha forest to indeed be the dominant systematic error on the Lyman-alpha forest auto-correlation. They additionally identify as significant terms the HCD absorber auto-correlation and a three-point correlation between two Lyman-alpha forest modes and an HCD absorber mode.

An approximate model for HCD absorbers is used in the measurement of the 3D Lyman-alpha forest correlation function with BOSS DR12 spectra (Bautista et al. 2017) and the cross-correlation with the quasar distribution (du Mas des Bourboux et al. 2017). It is (the Fourier transform of) a biased linear power spectrum, with separate bias and redshift space distortion parameters for HCD absorbers, convolved with a top-hat filter in real space (i.e. a sinc function in Fourier space) to approximate the profiles of HCD absorbers (hereafter the ‘BOSS model’). The large-scale bias of dark matter haloes hosting DLAs can be constrained through the cross-correlation of DLAs in spectra with the Lyman-alpha forest using BOSS spectra (Font-Ribera et al. 2012; Pérez-Ràfols et al. 2018). This halo bias can then be related to the absorber flux transmission bias.

In this study, we measure the effect of HCD absorbers on correlations in the 3D Lyman-alpha forest using the 3D flux power spectrum in cosmological hydrodynamical simulations for the first time. We use simulation boxes from the Illustris project (Vogelsberger et al. 2014a; Nelson et al. 2015), which have been shown to reproduce the observed column density distribution function (CDDF) and spatial clustering of HCD absorbers at the 95 per cent confidence level (Vogelsberger et al. 2014a; Bird et al. 2014). We measure the full anisotropic effect as a function of column density and redshift. We then consider how well the linear Voigt model characterizes our results and identify the regimes where this simple model breaks down. We also compare this model to the approximate BOSS model

discussed above. Our results will improve the robustness of modelling HCD absorbers and hence improve cosmological inference for future Lyman-alpha forest surveys (e.g. eBOSS/DESI).

We briefly explain the theory of the models that we consider for the Lyman-alpha forest and HCD absorbers in Section 2. In Section 3, our methodology in measuring the 3D flux power spectrum from hydrodynamical simulations and our modelling procedure are explained. We present our main results in Section 4. These results are discussed and compared to previous work in Section 5, and in Section 6, we draw our conclusions.

## 2 THEORY

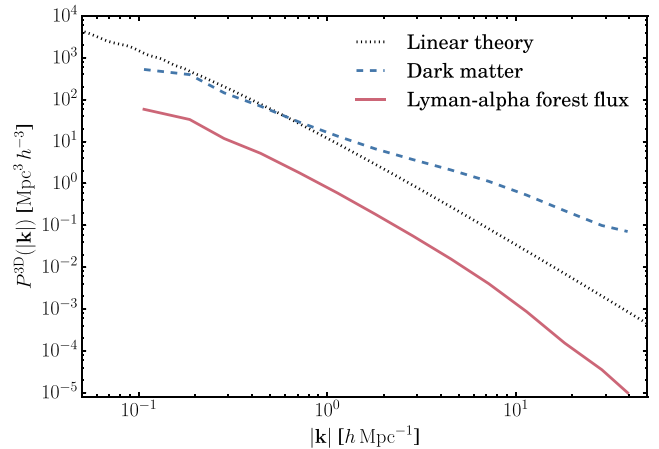
### 2.1 Lyman-alpha forest

Fluctuations in the transmitted flux of the Lyman-alpha forest are given as  $\delta_{\text{Forest}}(\mathbf{x}) = \frac{\mathcal{F}_{\text{Forest}}(\mathbf{x})}{\langle \mathcal{F}_{\text{Forest}} \rangle} - 1$ ,<sup>1</sup> where the transmitted flux  $\mathcal{F}_{\text{Forest}} = e^{-\tau_{\text{Forest}}}$ ,  $\tau_{\text{Forest}}$  is the optical depth and  $\langle \mathcal{F}_{\text{Forest}} \rangle$  is the average flux over all spectral pixels. We then follow the standard treatment of the Lyman-alpha forest on large scales and model these fluctuations as a biased tracer of the underlying matter density fluctuation field with redshift-space distortions (by analogy with other tracers of the matter distribution like galaxies or galaxy clusters; Kaiser 1984, 1987). It therefore follows that the 3D Lyman-alpha forest flux power spectrum can be modelled as

$$P_{\text{Forest}}^{\text{3D}}(|\mathbf{k}|, \mu, z) = b_{\text{Forest}}^2 (1 + \beta_{\text{Forest}} \mu^2)^2 P_{\text{Linear}}^{\text{3D}}(|\mathbf{k}|, z) D_{\text{NL}}(|\mathbf{k}|, \mu), \quad (1)$$

where (following McDonald 2003; Arinyo-i-Prats et al. 2015) we introduce a parametric function  $D_{\text{NL}}(|\mathbf{k}|, \mu)$  to characterize deviations from linear theory due to non-linear effects;  $P_{\text{Linear}}^{\text{3D}}(|\mathbf{k}|, z)$  is the linear theory matter power spectrum;  $b_{\text{Forest}}$  is the (linear) bias parameter of the Lyman-alpha forest; and  $\beta_{\text{Forest}}$  is its redshift space distortion parameter. For the wavevector  $\mathbf{k}$  (conjugate to  $\mathbf{x}$ ), we use a spherical coordinate system with its zenith direction along the line of sight such that power spectra are functions of  $|\mathbf{k}|$  and  $\mu$ , which is the cosine of the angle between the wavevector and the line of sight.  $P_{\text{Forest}}^{\text{3D}}$  is a function of redshift  $z$  and, in general, so are  $b_{\text{Forest}}$  and  $\beta_{\text{Forest}}$ . Constraints on the redshift evolution of the Lyman-alpha forest power spectrum largely come at present from the 1D flux power spectrum (although there was some analysis of the redshift evolution of the 3D power spectrum in Slosar et al. 2011). It is currently assumed (e.g. Bautista et al. 2017) that  $b_{\text{Forest}} \propto (1+z)^\gamma$ , where  $\gamma = 2.9$ , and that  $\beta_{\text{Forest}}$  does not depend on redshift such that  $b_{\text{Forest}}^2 P_{\text{Linear}}^{\text{3D}} \propto (1+z)^{3.8}$ , roughly matching the evolution observed in the 1D flux power spectrum (e.g. Palanque-Delabrouille et al. 2013). These assumptions are broadly supported by results from hydrodynamical simulations, although  $\beta_{\text{Forest}}$  is found to decrease with increasing redshift (e.g. Arinyo-i-Prats et al. 2015).

The bias parameters of the Lyman-alpha forest differ from the biases of individual sources like galaxies or haloes. For the latter, denser regions of matter contain a higher density of galaxies and haloes and so their bias parameters are positive. For the forest, denser regions of matter will have less transmitted flux (due to increased absorption by H I gas) and so  $b_{\text{Forest}}$  is negative. Fig. 1



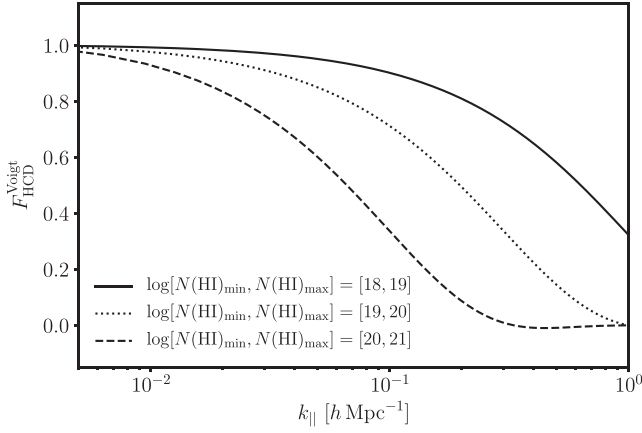
**Figure 1.** A comparison of 3D power spectra (averaged over all angles) as predicted by linear theory, measured from dark matter particles in a hydrodynamical simulation and measured from the transmission flux of the Lyman-alpha forest in redshift space from mock spectra generated from the same simulation. Although the Lyman-alpha forest is a biased tracer of linear theory, it remains linear to much smaller scales than other probes, including the power spectrum of dark matter, which is affected by non-linear gravitational evolution from scales larger than  $1 h \text{ Mpc}^{-1}$ . The simulation used is a  $(75 \text{ Mpc } h^{-1})^3$  box at redshift  $z = 2.44$  from the Illustris project (Vogelsberger et al. 2014a).

compares the 3D Lyman-alpha forest flux power spectrum (averaged over all angles) as measured in redshift space from our simulation (see Section 3.1) to the linear theory matter power spectrum. The flux power spectrum appears to be a scaled version of the theory power spectrum, remaining so too much smaller scales than, for example, the dark matter power spectrum, which is strongly affected by non-linear gravitational evolution for  $|\mathbf{k}| > 1 h \text{ Mpc}^{-1}$ . The deviation on small scales from linear theory is parametrized by the function  $D_{\text{NL}}$ , which is calibrated from hydrodynamical simulations. This function allows for the isotropic growth in power due to non-linear growth, isotropic suppression by pressure on very small scales, and suppression by non-linear peculiar velocities and temperature towards the line of sight. For consistency with Bautista et al. (2017), we use the fitting function of McDonald (2003) with the parameter values given in the first row of their table 1. The parameters of this function have not been measured from data or previous simulations at the higher redshift that we consider ( $z = 3.49$ ), and we are not able to do so with our simulations due to insufficient constraining power. Therefore, for simplicity, we use the low-redshift parameter values when modelling the high-redshift setting. Arinyo-i-Prats et al. (2015), in any case, found the shape of their simplest non-linear fitting function to evolve weakly in the redshifts they consider ( $2.2 \leq z \leq 3$ ).

### 2.2 High column density absorbers

We follow McQuinn & White (2011), Font-Ribera & Miralda-Escudé (2012), and Bautista et al. (2017) in modelling the 3D correlations of HCD absorbers on large scales. This is a linear model of HCD absorbers as point objects (i.e. without damping wings; akin to equation (1), convolved with the profiles of the wings). Therefore, the 3D flux power spectrum of a set of HCD absorbers with column densities in the interval  $[N(\text{H I})_{\text{min}}, N(\text{H I})_{\text{max}}]$  is given

<sup>1</sup> For the 3D comoving spatial coordinate  $\mathbf{x}$ , the line-of-sight component  $x_{\parallel}$  is transformed from the line-of-sight velocity space of spectra by the Hubble law.



**Figure 2.**  $F_{\text{HCD}}^{\text{Voigt}}(k_{\parallel}, z)$  (see equation 3) evaluated for the H I CDDF [ $f(N(\text{H I}), z)$ ] as measured in our simulation box (Illustris-1; Vogelsberger et al. 2014a) at  $z = 2.44$ . The units of  $N(\text{H I})$  are atoms  $\text{cm}^{-2}$ .

as

$$P_{\text{HCD}}^{\text{3D}}(|\mathbf{k}|, \mu, z) = b_{\text{HCD}}^2(1 + \beta_{\text{HCD}}\mu^2)^2 P_{\text{Linear}}^{\text{3D}}(|\mathbf{k}|, z) F_{\text{HCD}}^2(k_{\parallel}, z), \quad (2)$$

where  $b_{\text{HCD}}$  and  $\beta_{\text{HCD}}$  are the bias and redshift space distortion parameters of the absorption caused by HCD absorbers (these will in general depend on redshift  $z$ ).  $F_{\text{HCD}}$  is a function of the line-of-sight wavenumber  $k_{\parallel} = |\mathbf{k}|\mu$ , since it is caused by the absorption profiles of HCD absorbers which only manifest along the line of sight:

$$F_{\text{HCD}}^{\text{Voigt}}(k_{\parallel}, z) = \int_{N(\text{H I})_{\text{min}}}^{N(\text{H I})_{\text{max}}} dN(\text{H I}) f(N(\text{H I}), z) V(k_{\parallel}, N(\text{H I})). \quad (3)$$

Here,  $V(k_{\parallel}, N(\text{H I}))$  is the Fourier transform of the HCD absorbers' wing profiles as they manifest in the flux fluctuation field and  $f(N(\text{H I}), z)$  is the CDDF. The model we consider in this study uses the profile of HCD absorbers, which is a Voigt function in optical depth (see e.g. Appendix A of Rogers et al. 2018 for the full expression), the convolution of a Gaussian profile (caused by Doppler broadening), and a Lorentzian profile (caused by natural or collisional broadening). Fig. 2 shows the shape of  $F_{\text{HCD}}^{\text{Voigt}}(k_{\parallel}, z)$  for some representative values of  $[N(\text{H I})_{\text{min}}, N(\text{H I})_{\text{max}}]$ . It shows that HCD absorbers of lower column density, which have narrower wings, have their effect on the power spectrum on smaller scales. We will also consider the approximation made by the BOSS Collaboration (Bautista et al. 2017; du Mas des Bourboux et al. 2017), where the absorption profiles of HCD absorbers are modelled as top-hat filters (BOSS model):

$$F_{\text{HCD}}^{\text{BOSS}}(k_{\parallel}, z) = \frac{\sin(L_{\text{HCD}}k_{\parallel})}{L_{\text{HCD}}k_{\parallel}}, \quad (4)$$

where  $L_{\text{HCD}}$  is a free parameter setting the effective width of these filters.

By combining equations (1) and (2) (and additionally remembering the cross-correlation between the Lyman-alpha forest and HCD absorber fields), the 3D flux power spectrum for the Lyman-alpha forest contaminated by a set of HCD absorbers is given as

$$P_{\text{Contaminated}}^{\text{3D}}(|\mathbf{k}|, \mu, z) = P_{\text{Linear}}^{\text{3D}}(|\mathbf{k}|, z) \tilde{b}_{\text{Forest}}^2 D_{\text{NL}}(|\mathbf{k}|, \mu) + 2\tilde{b}_{\text{Forest}}\tilde{b}_{\text{HCD}} + \tilde{b}_{\text{HCD}}^2, \quad (5)$$

where  $\tilde{b}_{\text{Forest}} = b_{\text{Forest}}(1 + \beta_{\text{Forest}}\mu^2)$  and  $\tilde{b}_{\text{HCD}} = b_{\text{HCD}}(1 + \beta_{\text{HCD}}\mu^2)F_{\text{HCD}}(k_{\parallel}, z)$ . If there was uncertainty in the CDDF of a given sample of spectra, it will be preferable to sub-divide the column density integrals evaluated in the calculation of  $F_{\text{HCD}}$  in equation (3) and allow for extra terms in equation (5), with bias parameters ( $\tilde{b}_{\text{HCD},i}$ ) for the  $N$  categories of HCD absorbers

$$P_{\text{Contaminated}}^{\text{3D}}(|\mathbf{k}|, \mu, z) = P_{\text{Linear}}^{\text{3D}}(|\mathbf{k}|, z) \left[ \tilde{b}_{\text{Forest}}^2 D_{\text{NL}}(|\mathbf{k}|, \mu) + \sum_{i=1}^N \left( 2\tilde{b}_{\text{Forest}}\tilde{b}_{\text{HCD},i} + \sum_{j=1}^N \tilde{b}_{\text{HCD},i}\tilde{b}_{\text{HCD},j} \right) \right]. \quad (6)$$

We also mention two possible additions that could be made to this model. First, the model in equation (5) does not consider any non-linear evolution in the clustering of HCD absorbers.<sup>2</sup> Secondly, as noted in Font-Ribera & Miralda-Escudé (2012), in the two-point function of the total contaminated flux, there will arise three- and four-point functions of the Lyman-alpha forest and HCD absorber fluctuations. This is because the forest and HCD absorption terms are multiplied:  $\langle \mathcal{F}_{\text{Total}} \rangle (1 + \delta_{\text{Total}}) = \langle \mathcal{F}_{\text{Forest}} \rangle (1 + \delta_{\text{Forest}}) \langle \mathcal{F}_{\text{HCD}} \rangle (1 + \delta_{\text{HCD}})$ , where  $\mathcal{F}_{\text{HCD}}$  is the flux transmitted by HCD absorbers and  $\delta_{\text{HCD}} = \mathcal{F}_{\text{HCD}} / \langle \mathcal{F}_{\text{HCD}} \rangle - 1$ . It follows that in the total flux power spectrum (equation 5), there will be three- and four-point correlations involving  $\delta_{\text{Forest}}$  and  $\delta_{\text{HCD}}$ . The model presented in equation (5) only accounts for the leading two-point correlations; Font-Ribera & Miralda-Escudé (2012), however, found that the three-point term  $\langle \delta_{\text{Forest}}(\mathbf{x}_1)\delta_{\text{HCD}}(\mathbf{x}_1)\delta_{\text{Forest}}(\mathbf{x}_2) \rangle$  is an important term on smaller scales (separations  $r < 40 \text{ Mpc } h^{-1}$ ). We discuss the possible impact of this additional term in Section 5.

### 3 METHOD

We first outline the method we have used and then explain the steps in more detail in the following subsections (Section 3.1 to 3.4).

(1) We use a cosmological hydrodynamical simulation from the Illustris project (Vogelsberger et al. 2014a; Nelson et al. 2015) and generate mock spectra on a grid (562 500 in total, each at a velocity resolution of  $10 \text{ km s}^{-1}$  and with a typical length of  $\sim 8000 \text{ km s}^{-1}$ ). We calculate these at two redshift slices ( $z = [2.44, 3.49]$ ; see Section 3.1).

(2) The mock spectra we generate contain absorption from the Lyman-alpha forest and HCD absorbers. For our analysis, it is useful to have a set of spectra containing only the Lyman-alpha forest [still forming a regular grid to allow the use of fast Fourier transforms (FFTs)]. To achieve this, we replace spectra contaminated by HCD absorbers by a nearby spectrum containing only the forest. Furthermore, we are able to construct boxes of spectra containing only the Lyman-alpha forest and a particular category of HCD absorber (i.e. restricted to a particular column density interval) by replacing back the original spectra containing only that category of contamination. The details of this HCD ‘dodging’ procedure are explained in Section 3.2.

<sup>2</sup> Indeed, the full ‘BOSS model’ as used by Bautista et al. (2017) and du Mas des Bourboux et al. (2017) multiplies the last two terms in equation (5) by  $D_{\text{NL}}$ , the non-linear function calibrated by simulations of the Lyman-alpha forest only. We do not, in the first instance, include this correction to the linear Voigt model.



(3) For each box of spectra that we generate, we measure the 3D flux power spectrum using an FFT (see Section 3.3).

(4) Using these measurements of 3D flux power spectra, we fit the proposed model (equation 5) using a Markov chain Monte Carlo (MCMC) method (see Section 3.4 and Appendix B).

### 3.1 Hydrodynamical simulations and mock spectra

We use snapshots from the highest resolution cosmological hydrodynamical simulation of the original Illustris project (Vogelsberger et al. 2014a; Nelson et al. 2015, Illustris-1<sup>3</sup>). The simulation adopts the following cosmological parameters:  $\Omega_m = 0.2726$ ,  $\Omega_\Lambda = 0.7274$ ,  $\Omega_b = 0.0456$ ,  $\sigma_8 = 0.809$ ,  $n_s = 0.963$ , and  $H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}$ , where  $h = 0.704$  (Vogelsberger et al. 2014b). The box has a comoving volume of  $(106.5 \text{ Mpc})^3$ , and we consider snapshots at redshifts  $z = 2.44$  and  $3.49$ . Illustris-1 has  $1820^3$  dark matter particles, each of mass  $6.3 \times 10^6 M_\odot$ ; the gas particle masses are each  $1.3 \times 10^6 M_\odot$ . The simulations are in broad agreement with observations of the H I CDDF (Vogelsberger et al. 2014a), the clustering of DLA haloes (Bird et al. 2014), and the kinematics of HCD absorbers (Bird et al. 2015). For a summary of the relevant physics in the simulation and comparisons to observations, see Rogers et al. (2018). This broad agreement with relevant observations justifies the use of this simulation in building and testing models of HCD absorbers for use in future data analyses. This is further supported by the combination of large box size and high resolution. This allows the galaxy formation physics that is essential for correctly modelling HCD absorbers (which are generally associated with high-redshift galaxies) to be rendered in high resolution, while simultaneously allowing measurement of the large-scale behaviour. In particular, the large box size allows the generation of long mock spectra (see below), which can span the widths of large damping wings. This was not achievable with previous generations of simulations. A comparable simulation suite comes from the EAGLE project (Schaye et al. 2015), although their largest box is slightly smaller and has slightly lower mass resolution than Illustris-1.

For each snapshot, we generate mock spectra (using the `fake_spectra` code; Bird 2017) containing only the Lyman-alpha absorption line, on a square grid of  $750^2 = 562\,500$  spectra, giving a spacing of  $142 \text{ kpc}$  between neighbouring spectra along each axis of the simulation box. Each spectrum extends the full length of the simulation box with periodic boundary conditions, giving a size in velocity space of  $7501$  and  $8420 \text{ km s}^{-1}$ , respectively, at  $z = 2.44$  and  $3.49$ . We measure the optical depth  $\tau$  in velocity bins of size  $10 \text{ km s}^{-1}$  along the spectrum. We then calculate the transmitted flux  $\mathcal{F} = e^{-\tau}$ . We convolve our spectra with a Gaussian kernel of  $\text{FWHM} = 8 \text{ km s}^{-1}$ , setting the simulated spectrographic resolution.

### 3.2 Dodging high column density absorbers

We associate with each mock spectrum that is contaminated by HCD absorbers a nearby spectrum containing only Lyman-alpha forest absorption. Indeed, if the transverse distance necessary to ‘dodge’ the contaminating HCD absorber is small (as is expected considering the physical sizes of HCD absorbers, Krogager et al. 2012), the large-scale cosmological modes (i.e. the Lyman-alpha

forest modes) in the replacement spectrum should be identical to the original spectrum and the difference will be the HCD absorber modes only. The ‘dodging’ procedure iteratively proposes a nearby replacement spectrum until one is found with no HCD absorber contamination (i.e. there are no column densities, integrated over  $100 \text{ km s}^{-1}$ ,<sup>4</sup> exceeding  $1.6 \times 10^{17} \text{ atoms cm}^{-2}$ , the threshold for HCD absorbers). It searches for replacement spectra by successively generating spectra further away in a transverse direction from the original spectrum in steps of  $10 \text{ kpc } h^{-1}$  until a suitable spectrum is found. In this way, we are able to generate a box of spectra containing only the Lyman-alpha forest.

We are also able to generate boxes of spectra containing Lyman-alpha forest and HCD absorbers of a certain category (i.e. column densities in a certain interval) by replacing back original spectra containing this particular category. We categorize spectra according to the maximum column density (again integrated over  $100 \text{ km s}^{-1}$ ) in each spectrum; there may be less dense HCD absorbers in each category, but their effect will be sub-dominant since their damping wings are narrower.

Having generated these new boxes of spectra, we compute FFTs (Section 3.3), ignoring the transverse dodging distances and assuming that the dodged spectra lie on the original grid. Since the dodging distances are, in general, small (we find only  $\sim 1$  percent to be  $> 500 \text{ kpc } h^{-1}$ ; see Appendix A), the error associated with this approximation is restricted to small scales, i.e. large  $|\mathbf{k}|$ . We conduct an analysis of the error that arises from the irregular grid resulting from the dodging distances in Appendix A. Following these tests, we study only scales  $|\mathbf{k}| < |\mathbf{k}|_{\text{max}}$ , where  $|\mathbf{k}|_{\text{max}} = 1 h \text{ Mpc}^{-1}$ ; at these small values of  $|\mathbf{k}|$ , the dodging error is negligible compared to the effect of HCD absorbers that we wish to measure (see Fig. 4).

### 3.3 Three-dimensional flux power spectrum

We measure the 3D flux power spectrum at each redshift slice for our Lyman-alpha forest box of spectra ( $P_{\text{Forest}}^{\text{3D}}$ ) and for our contaminated boxes of spectra for a number of HCD absorber categories ( $P_{\text{Contaminated}}^{\text{3D}}$ ) the column density ranges of which we give in Table 1. We estimate the 3D flux power spectrum in bins of  $|\mathbf{k}|$  (15 bins) and  $\mu$  (4 bins),  $P_{\text{Flux},i}^{\text{3D}} = \frac{1}{N_i} \sum_n |\hat{\delta}_{\text{Flux}}(\mathbf{k}_n)|^2$ , where  $\mathbf{k}_n$  lie within a given  $(|\mathbf{k}|, \mu)$  bin and  $N_i$  are the number of modes in each bin  $i$ .  $\hat{\delta}_{\text{Flux}}(\mathbf{k})$  is the Fourier transform of the flux fluctuation field  $\delta_{\text{Flux}}(\mathbf{x}) = \frac{\mathcal{F}(\mathbf{x})}{\langle \mathcal{F} \rangle} - 1$ . Here, for the mean flux  $\langle \mathcal{F} \rangle$ , we always use the mean flux of the original box of spectra (with no dodging) so that our modelling assumption that the flux fluctuations can be sub-divided into different absorber categories  $\delta_{\text{Total}} = \sum_i \delta_i$  holds true, where  $i$  indexes the different absorber categories. We use the convention of absorbing the  $(2\pi)^3$  into the conjugate variable, i.e. we define the Fourier transform as  $\delta(k) = \int \delta(x) e^{-ikx} dx$ .

### 3.4 Modelling and Markov chain Monte Carlo sampling

We optimize the parameters of our model (the Lyman-alpha forest and HCD contamination biases and redshift-space distortion parameters) using MCMC sampling. We use MCMC sampling in order to

<sup>4</sup> This is the same integration length as we used in our 1D flux power spectrum analysis in Rogers et al. (2018) and it amounts to 10 neighbouring bins or a comoving length much larger than the most extensive HCD absorbers (Krogager et al. 2012). In Rogers et al. (2018), we also tested our sensitivity to the size of this integration length and found it made negligible difference to power spectrum estimates.

<sup>3</sup> The simulation we use is publically available at <http://www.illustris-project.org/data>.

**Table 1.** The neutral hydrogen ( $\text{H I}$ ) column density limits  $[N(\text{H I})_{\min}, N(\text{H I})_{\max}]$  that define the categories of absorbing systems used in this work. The columns on the right show the percentage of spectra (at each redshift  $z$  that is considered) in our  $(106.5 \text{ Mpc})^3$  simulation box (Vogelsberger et al. 2014a; Nelson et al. 2015, Illustris-1), where the highest density system belongs to a given category.

Absorber category	$N(\text{H I})_{\min}$ [atoms $\text{cm}^{-2}$ ]	$N(\text{H I})_{\max}$	per cent of spectra in box at	
			$z = 2.44$	$z = 3.49$
Lyman- $\alpha$ forest	0	$1.6 \times 10^{17}$	69.6	45.7
LLS	$1.6 \times 10^{17}$	$1 \times 10^{19}$	14.9	27.0
Sub-DLA	$1 \times 10^{19}$	$2 \times 10^{20}$	8.1	14.3
Small DLA	$2 \times 10^{20}$	$1 \times 10^{21}$	4.1	7.8
Large DLA	$1 \times 10^{21}$	$\infty$	3.3	5.2

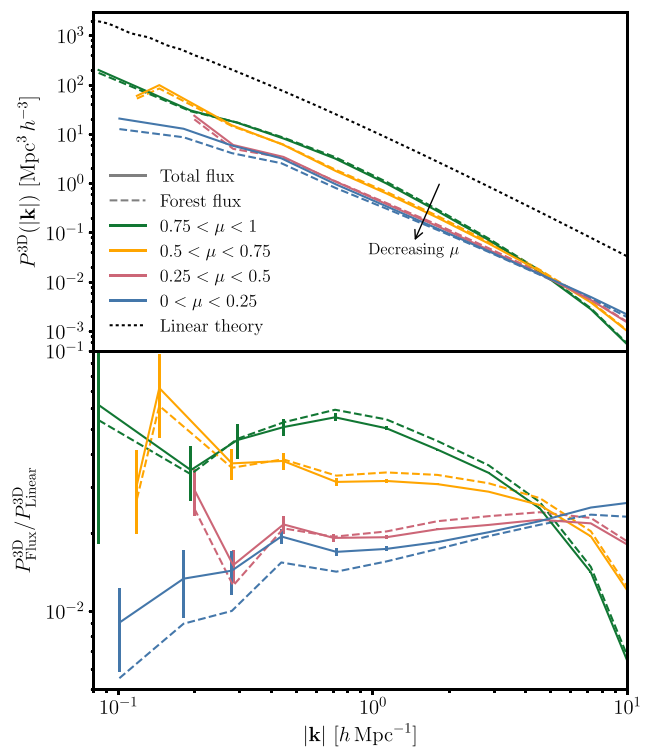
estimate parameter uncertainties and to understand parameter degeneracies. Our data vectors consist of the flux power spectra of a contaminated set of spectra and an uncontaminated set (containing only Lyman-alpha forest). We model each data point as chi-square distributed (as is explained in Appendix B) with an estimated variance of  $2(P_{\text{True},i}^{\text{3D}})^2/N_i$ , where  $P_{\text{True},i}^{\text{3D}}$  is the true (ensemble) value of the 3D flux power spectrum in bin  $i$  and  $N_i$  is the number of modes per bin. These variances form the elements of our diagonal covariance matrices. The full details of the construction of our likelihood function and prior probability distributions are given in Appendix B. The results are shown in Section 4.2.

## 4 RESULTS

### 4.1 Measuring the effect of HCD absorbers

Fig. 3 shows the measured 3D flux power spectra as a function of scale  $|\mathbf{k}|$  and the cosine of the angle away from the line of sight  $\mu$ ; i.e.  $\mu = 1$  is along the line of sight and  $\mu = 0$  is transverse to the line of sight. Anisotropic behaviour arises due to linear redshift-space distortions on larger scales, enhancing power towards the line of sight. Non-linear effects on smaller scales suppress power along the line of sight due to non-linear peculiar velocities and thermal broadening of absorption lines. The non-linear effects are more manifest in the bottom panel, where ratios to the linear matter power spectrum are shown. On the largest scales, these ratios should tend towards constant values (i.e.  $b^2(1 + \beta\mu^2)^2$  for the linear models presented in Section 2). However, we are only able to probe a small number of these large-scale modes in our  $75 \text{ Mpc } h^{-1}$  simulation box and so our measurement of large-scale bias has a large variance. The isotropic enhancement of power due to non-linear collapse of structure is broadly observable on larger scales ( $|\mathbf{k}| \sim 0.03 \text{ h Mpc}^{-1}$ ), although this trend is also obscured by the large variance on large scales. The anisotropic suppression of power towards the line of sight mentioned above is clearly observable for scales  $|\mathbf{k}| > 1 \text{ h Mpc}^{-1}$ , leading to the characteristic cross-over in the curves on small scales (McDonald 2003; Arinyo-i-Prats et al. 2015).

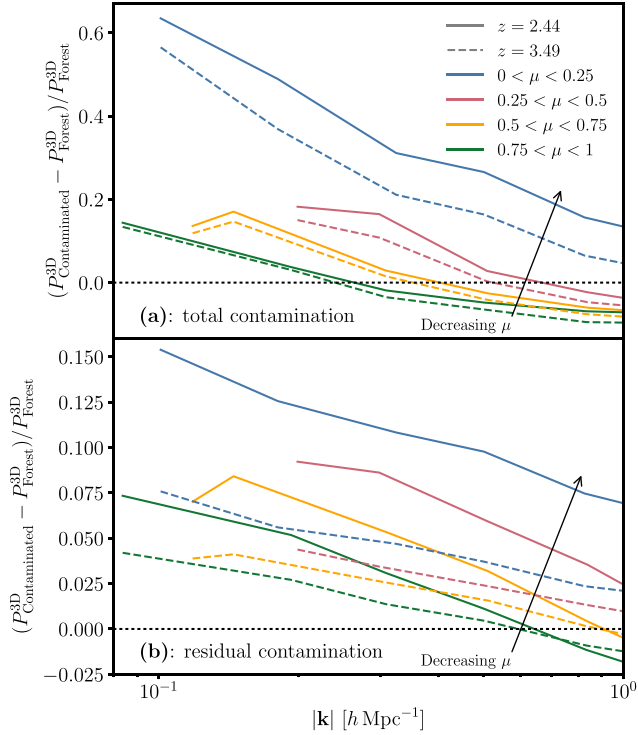
Fig. 3 also compares the 3D flux power spectra of contaminated and uncontaminated Lyman-alpha forest absorption (solid and dashed lines, respectively; see Section 3.2 for more details about how a box of spectra without HCD absorber contamination is constructed by the dodging technique). The contamination by HCD absorbers adds power in some regimes (especially in the transverse direction) and suppresses power in others (especially on smaller scales towards the line of sight). Following the tests of the error caused by the dodging procedure in forming the Lyman-alpha forest box of spectra (see Appendix A and Section 3.2), we cut our



**Figure 3.** Above: the 3D power spectra of the total flux from the Lyman-alpha forest and HCD absorbers (solid lines); and of the flux from the Lyman-alpha forest only (dashed lines); and the total linear theory matter power spectrum. For the flux power spectra, we show the anisotropic behaviour as a function of  $\mu$ . Below: the flux power spectra in ratio to the linear power spectrum. Flux measurements are made from a simulation box at redshift  $z = 2.44$ .

data vectors at  $|\mathbf{k}|_{\max} = 1$ , and we will throw away smaller scales in our following analysis. The forest flux power spectra shown in Fig. 3 (dashed lines) do not quantify the additional systematic and statistical error arising from the dodging (which is only significant for  $|\mathbf{k}| > 1 \text{ h Mpc}^{-1}$ ).

Fig. 4 shows the fractional effect of HCD absorber contamination on the 3D Lyman-alpha forest flux power spectrum. The fractional effect of the full ensemble of HCD absorbers [panel (a)] can be as large as a 60 per cent correction to  $P^{\text{3D}}(|\mathbf{k}|)$  at  $|\mathbf{k}| = 0.1 \text{ h Mpc}^{-1}$  in the transverse direction. The fractional effect is smaller at higher redshift because the Lyman-alpha forest power spectrum (in the denominator) has a larger amplitude (since neutral hydrogen is more abundant and so there is stronger Lyman-alpha absorption). There is a larger fractional effect in the transverse direction



**Figure 4.** The change in the flux power spectrum from contamination of quasar spectra by HCD absorbers, as a fraction of the Lyman-alpha forest power spectrum. (a) above: the effect of the total contamination from all HCD absorbers in our simulation box; (b) below: the effect of a mock residual contamination after the largest HCD absorbers have been ‘clipped’ from quasar spectra (i.e. only LLS and sub-DLAs remaining).

driven also by the Lyman-alpha forest power spectrum, which has less power in this direction due to redshift-space distortions. The scale-dependence in Fig. 4 is partly driven by the non-linear effects in the Lyman-alpha forest power spectrum discussed above; in particular, the Lyman-alpha forest power spectrum is boosted on small scales due to non-linear growth, and so the fractional effect decreases.

The bottom panel Fig. 4 (b) shows the equivalent effects but for a mock residual contamination of HCD absorbers after the largest HCD absorbers have been ‘clipped’ out (i.e. only LLS and sub-DLAs remaining). The same trends are observed as above, but the overall amplitude is smaller since the largest damping wings have been removed. None the less, the effect at  $|\mathbf{k}| = 0.1 h \text{ Mpc}^{-1}$  in the transverse direction still constitutes a 15 per cent correction; it is therefore necessary to model this effect for robust cosmological inference from the Lyman-alpha forest (see Section 4.2).

#### 4.2 Modelling the effect of HCD absorbers

We show the maximum posterior values of the linear Voigt model and the measurements made in our simulations in Fig. 5. To emphasize the effect of HCD absorbers, we show the part of the model for the auto-correlations of HCD absorbers and their cross-correlation with the Lyman-alpha forest, i.e. the last two terms in equation (5) ( $2\tilde{b}_{\text{Forest}}\tilde{b}_{\text{HCD}} + \tilde{b}_{\text{HCD}}^2$ ). We compare this to the difference between the flux power spectra of the contaminated and uncontaminated boxes of spectra, in ratio to the linear theory matter power spectrum ( $(P_{\text{Contaminated}}^{\text{3D}} - P_{\text{Forest}}^{\text{3D}})/P_{\text{Linear}}^{\text{3D}}$ ). We plot the results as a function of column density (by showing the effect for different HCD ab-

sorber categories from the top to bottom) and as a function of redshift (from the left- to right-hand panel). The error bars scale appropriately with the number of modes in each power spectrum bin ( $= d_i \sqrt{2/N_i}$ ), where  $d_i$  is the data point value but do not capture the full likelihood function, the details of which are given in Appendix B.

A measure of the goodness of fit is the values of the reduced chi-squared statistic: from the top to bottom, the left- to right-hand panel,  $\chi_{\text{red}}^2 =$  (a) 1.55; (b) 1.56; (c) 1.57; (d) 1.73; (e) 1.04; (f) 1.03; (g) 1.21; and (h) 1.24, all indicating a good fit for the linear Voigt model for all the column densities and redshifts we have considered, excluding certain regimes as explained below. The number of degrees of freedom is (a), (b), (e), (f): 34; (c), (g): 26; (d), (h): 22. The linear Voigt model is discrepant with the simulation results for a small part of the data space ( $|\mathbf{k}| \gtrsim 0.4 h \text{ Mpc}^{-1}$  for small DLAs and  $|\mathbf{k}| \gtrsim 0.25 h \text{ Mpc}^{-1}$  for large DLAs); these exceptions are indicated by the dotted lines in Fig. 5 and are restricted to small scales (particularly towards the line of sight) for the largest HCD absorbers (small and large DLAs). Indeed, we exclude these parts of the data vector for the contaminated 3D flux power spectra in our parameter inference; we discuss the implications of this small-scale discrepancy for the highest column density absorbers in Section 5. Part of the discrepancy between simulation and model in Fig. 5 is also driven by the range of  $\mu$  values within each  $\mu$  bin.

For completeness, we quote the maximum (marginalized) posterior values and  $1\sigma$  credible intervals of the (linear) bias parameters of the Lyman-alpha forest<sup>5</sup> at  $z = 2.44$ :

$$b_{\text{Forest}}(1 + \beta_{\text{Forest}}) = -0.270 \pm 0.004; \quad \beta_{\text{Forest}} = 1.722 \pm 0.072$$

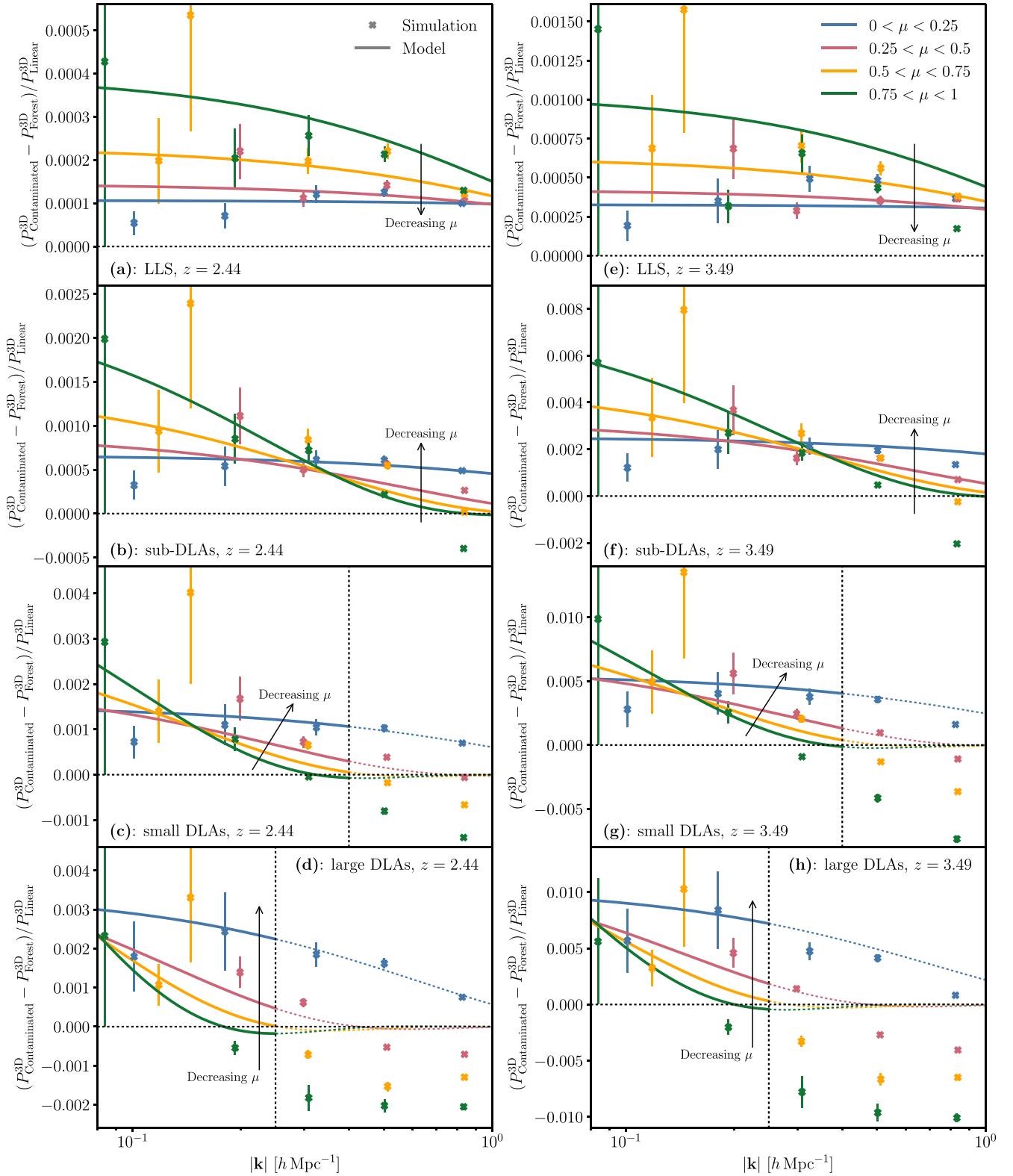
and at  $z = 3.49$ :

$$b_{\text{Forest}}(1 + \beta_{\text{Forest}}) = -0.511 \pm 0.006; \quad \beta_{\text{Forest}} = 1.249 \pm 0.043.$$

The posterior on  $\beta_{\text{Forest}}$  at  $z = 2.44$  is in  $1\sigma$  agreement with the best-fitting value from BOSS DR12 spectra (Bautista et al. 2017),  $\beta_{\text{Forest}}^{\text{BOSS}} = 1.663 \pm 0.085$  at a central redshift of  $z = 2.3$ . However, the posterior on  $b_{\text{Forest}}(1 + \beta_{\text{Forest}})$  is lower than the value measured from data  $b_{\text{Forest}}^{\text{BOSS}}(1 + \beta_{\text{Forest}}^{\text{BOSS}}) = -0.325 \pm 0.004$  at  $z = 2.3$ ; this difference has been observed in other studies with hydrodynamical simulations (e.g. Arinyo-i-Prats et al. 2015). The redshift evolution in  $b_{\text{Forest}}$  observed in our simulations (modelled as  $b_{\text{Forest}} \propto (1+z)^\gamma$ ) implies  $\gamma = 3.1$ , roughly matching the value currently assumed in data analyses  $\gamma^{\text{BOSS}} = 2.9$  (see Section 2.1). We find that  $\beta_{\text{Forest}}$  decreases at higher redshift, also as observed in previous studies with simulations (Arinyo-i-Prats et al. 2015). In Appendix C, we test the sensitivity of our inference of the bias parameters of the Lyman-alpha forest to the smallest scale included in our analysis  $|\mathbf{k}|_{\text{max}}$ ; we find that our inferences are overall insensitive to this, suggesting that our results are robust to our modelling of non-linear effects. We also recover the same posterior distributions on the Lyman-alpha forest bias parameters for each of the HCD absorber categories of contaminated flux power spectra that we consider. These parameters also match those inferred from the 3D Lyman-alpha forest flux power spectrum only.

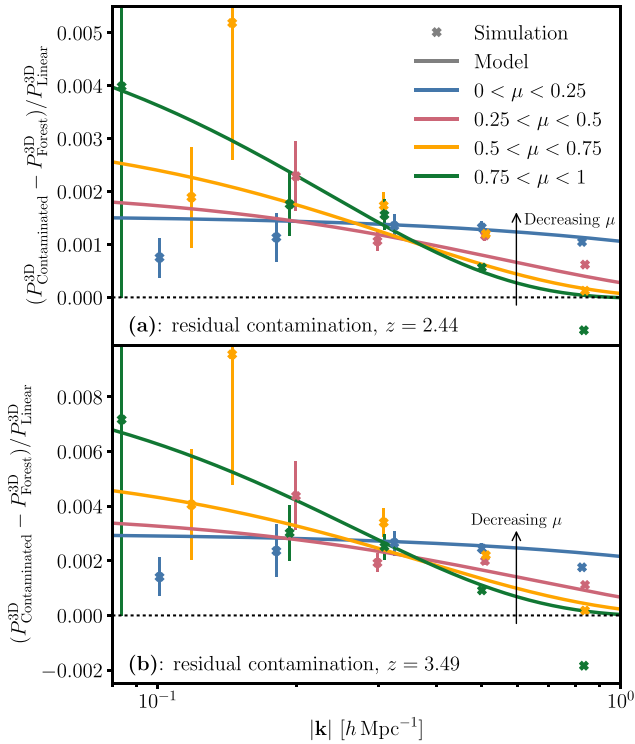
As discussed in Section 3.4, we place Gaussian priors on the bias parameters of the different categories of HCD absorber (but not the forest biases), which are otherwise poorly constrained, since the amplitude of their effect is sub-dominant to the Lyman-alpha forest

<sup>5</sup> Rather than  $b_{\text{Forest}}$ , following e.g. Slosar et al. (2011); Bautista et al. (2017), we sample the combination  $b_{\text{Forest}}(1 + \beta_{\text{Forest}})$ , which is less correlated with  $\beta_{\text{Forest}}$ .



**Figure 5.** The change in the flux power spectrum from contamination of quasar spectra by different categories of HCD absorbers, in ratio to the linear power spectrum. The points are measurements from our simulation boxes; error bars indicate the number of modes in each bin. The lines are maximum posterior values of our preferred model. *From the top to bottom panel, we show the effect of different categories of HCD absorbers; from the left to right-hand panel, we show the effect at different redshifts  $z$ .* The vertical dotted lines for the two largest HCD absorber categories indicate the smallest scale that we include in our data vector from our HCD-contaminated simulation boxes for those categories. Our preferred model does not correctly characterize the simulation results for these categories on smaller scales towards the line of sight. The dotted lines show an extrapolation of this model, highlighting the discrepancy.





**Figure 6.** As Fig. 5, but for a mock residual contamination after the largest HCD absorbers have been ‘clipped’ from quasar spectra (i.e. only LLS and sub-DLAs remaining). From the top to bottom panel, we show the effect at different redshifts  $z$ .

flux power spectrum. These prior distributions are returned almost exactly in the marginalized posteriors. The scale-dependence of the effect of HCD absorbers, meanwhile, is fully determined by the physics of their absorption profiles and the appropriate CDDF.

Fig. 6 compares the maximum posterior values of the linear Voigt model to the effect of a mock residual contamination of HCD absorbers on the 3D flux power spectrum at the two redshifts we consider. This mock residual contamination (of LLS and sub-DLAs) approximately matches the column densities assumed remaining in BOSS spectra (Bautista et al. 2017) after the largest damping wings have been removed. We find that the linear Voigt model is in statistical agreement with our simulation measurements ( $\chi^2_{\text{red}} =$  (a) 1.56; (b) 1.03). The number of degrees of freedom is (for both panels) 34. We highlight this configuration because, although Fig. 5 shows that there are some small scales towards the line of sight for the largest HCD absorbers where the simple linear Voigt model is not appropriate, these are the HCD absorber categories most efficiently removed in the ‘clipping’ process in data analysis.

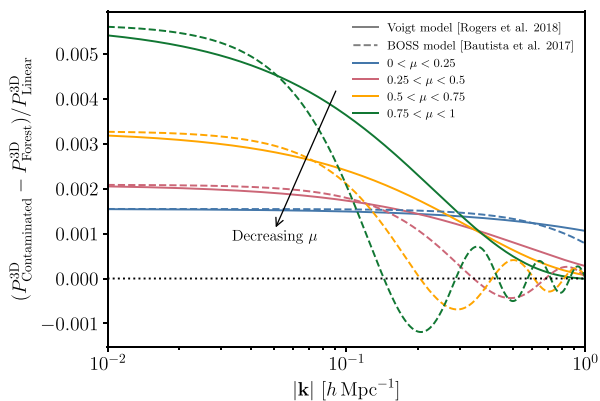
## 5 DISCUSSION

In Section 4.2, we showed the regimes in scale and column density where the simple linear Voigt model (see Section 2.2) can characterize the effect of HCD absorber contamination on correlations in the 3D Lyman-alpha forest. The linear Voigt model is arguably the simplest model that can be constructed to take account of the true absorption line profiles of HCD absorbers. It is a linear model with HCD absorbers as biased tracers of the matter density distribution with redshift-space distortions. This is then convolved with the Voigt profiles of HCD absorbers’ damping wings and integrated over the CDDF of the absorbers. The linear Voigt model manifests

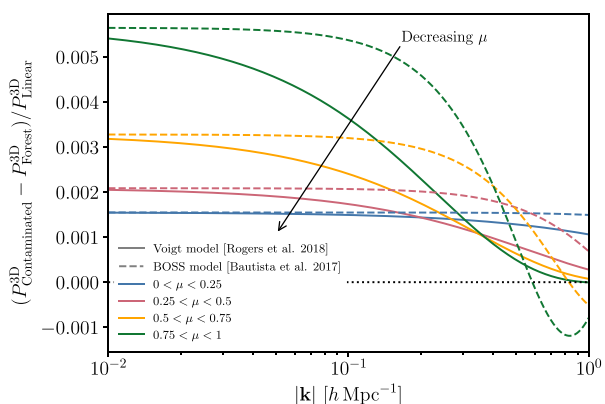
as a suppression in power due to damping wings that remove structure in the spectra. This effect is stronger towards the line of sight since this is the direction in which the wings appear. The scales at which the suppression starts (and the overall amplitude) are larger for more dense absorber categories since their wings are wider. The suppression is such that it overcomes the boost in power towards the line of sight on large scales due to redshift-space distortions; consequently, there is a characteristic cross-over in the different curves for each absorber category. The effect of HCD absorbers transverse to the line of sight is scale-independent, since there is no component of the damping wings in this direction (our lowest  $\mu$  bin does include some modes slightly away from the transverse direction). The amplitude of the effect increases with redshift, mainly because the cross-correlation with the Lyman-alpha forest is stronger (there being overall more absorption at higher redshift).

However, on small scales towards the line of sight for the largest HCD absorber categories ( $|\mathbf{k}| \gtrsim 0.4 h \text{ Mpc}^{-1}$  for small DLAs and  $|\mathbf{k}| \gtrsim 0.25 h \text{ Mpc}^{-1}$  for large DLAs), the linear Voigt model cannot characterize our simulation results. Power is suppressed towards the line of sight more strongly than our model allows such that there is less power than without the HCD absorbers. We consider two possible causes of this discrepancy with the linear Voigt model (as mentioned in Section 2.2). First, our model does not consider any non-linear clustering of the gas or haloes associated with HCD absorbers. A comprehensive model for the clustering of HCD absorbers should certainly account for this effect. However, we found no preference for a parametric form (akin to that used for the Lyman-alpha forest; McDonald 2003; Arinyo-i-Prats et al. 2015) that would improve the fit to our simulation results. This suggests that such a closed form cannot alone account for the discrepancy. Secondly, we consider the non-linear effect of a three-point correlation between a Lyman-alpha forest fluctuation and an HCD absorber fluctuation at the same position and a Lyman-alpha forest fluctuation at a second position (see Section 2.2). This term was shown by Font-Ribera & Miralda-Escudé (2012) to be at least as significant as the HCD absorber auto-correlation on small scales (separations  $r < 40 \text{ Mpc } h^{-1}$ ) and to have the correct (negative) sign to account for the additional suppression of power observed on small scales towards the line of sight for small and large DLAs. (This is because it is the correlation between three negatively biased tracers.) It is intuitively understood as the effect of the damping wings in masking regions of the Lyman-alpha forest and so suppressing auto-correlations in the Lyman-alpha forest that would otherwise occur on scales within the widths of individual wings. This effect will be stronger for more dense HCD absorbers since their damping wings are wider and so mask more of the Lyman-alpha forest; stronger on scales smaller than the widths of wings; and stronger towards the line of sight since this is the direction in which the masking occurs. This seems a qualitative match to the observed discrepancies with the linear Voigt model, but as yet there exists no simple model for this higher order effect, and we have not explicitly tested whether it can account for the observed discrepancies. As discussed above and in Section 4.2, the effect is restricted to the highest column densities, which are in any case mostly removed in the clipping pre-processing of spectra.

Fig. 7 compares the linear Voigt model as inferred from our simulations contaminated by a mock residual contamination of HCD absorbers (only LLS and sub-DLAs remaining, approximating the effect of clipping out the damping wings of more dense absorbers as is done with survey spectra), with the model used by the BOSS Collaboration (Bautista et al. 2017; du Mas des Bourboux et al. 2017) for the same effect. This ‘BOSS model’ approximates the



**Figure 7.** A comparison of the existing model as used by the BOSS Collaboration (Bautista et al. 2017; Pérez-Ràfols et al. 2018) and the linear Voigt model presented in this study for the additive effect of residual HCD absorbers (after the ‘clipping’ of the largest absorbers from quasar spectra) on the 3D flux power spectrum. For the linear Voigt model, we show the maximum posterior values as inferred from a mock residual contamination in our simulation box at  $z = 2.44$ . For the BOSS model, we rescale to match the bias and redshift-space distortions inferred in our box, but use the best-fitting value of the shape parameter as found in BOSS mock spectra with a residual contamination and data. The maximum posterior value of the BOSS model as inferred from our simulation gives unphysical results on scales larger than the size of our box (see Fig. 8).



**Figure 8.** As Fig. 7, but the BOSS model is shown for the maximum posterior value of the shape parameter as inferred from our simulation. This highlights how the BOSS model, in attempting to fit the small-scale behaviour in our box, gives a flat response on scales larger than the size of our box. This flat response suggests that there is no large-scale effect of HCD absorbers. This contradicts the physical linear Voigt model and the results from data and mock spectra from the BOSS Collaboration (Bautista et al. 2017; du Mas des Bourboux et al. 2017; Pérez-Ràfols et al. 2018).

damping wings as top-hats and so the effect on the Fourier space correlations (i.e. the flux power spectrum) is a sinc function (see Section 2.2). We rescale the BOSS model to have the same bias and redshift-space distortions as inferred in our simulation box (for a fair comparison to the linear Voigt model), but use the shape parameter as found in BOSS mock spectra with a residual contamination and in data ( $L_{\text{HCD}} = 24.341 \text{ Mpc } h^{-1}$ ).<sup>6</sup> We extrapolate the BOSS model to smaller scales than considered in their analysis, where the

minimum separations measured were  $r = 10 \text{ Mpc } h^{-1}$ . Although our inference on the linear Voigt model is only constrained by the scales accessible in our simulation box, we extrapolate this model to larger scales of relevance to a BAO analysis. Although we are not able to explicitly test the model on these larger scales, it is expected to correctly characterize the effect as it constitutes the physical expectation on large scales. We conclude from Fig. 7 that the BOSS model constitutes a good approximation for scales of relevance for a BAO analysis, but that on smaller scales, the linear Voigt model should be used in order to account for the effect of extended damping wings for a residual contamination of HCD absorbers. There is always a residual contamination because it is difficult to identify narrow damping wings amongst the superposed Lyman-alpha forest absorption lines and instrumental noise. This becomes harder for noisier quasar spectra. By definition, it is impossible to know the exact efficiency of detection algorithms on survey data, although this can be estimated on simulations. However, the model we have constructed (as given by equation 6) has the flexibility to marginalize over this uncertainty, by sub-dividing HCD absorbers into categories, each with their own set of parameters. Appropriate priors reflecting the expected down-weighting of higher column density systems can be constructed based on previous analyses and simulation testing; the details will be survey-specific.

## 6 CONCLUSIONS

We have measured the effect of contamination of quasar spectra by the damping wings of HCD absorbing regions of neutral hydrogen on correlations in the 3D Lyman-alpha forest. We accomplished this by measuring 3D flux power spectra from a cosmological hydrodynamical simulation (Illustris; Vogelsberger et al. 2014a; Nelson et al. 2015) as a function of the column density of the HCD absorber contamination and redshift. We found that, even after the largest damping wings have been removed (as performed by survey pipelines), that the effect of the residual contamination can be as large as a 15 per cent correction to the 3D Lyman-alpha forest flux power spectrum (at  $|k| = 0.1 \text{ h Mpc}^{-1}$ ). We found that the effect of this residual contamination can be characterized by a simple linear model (with bias and redshift-space distortions) convolved with the Voigt profiles of the damping wings and integrated over the CDDF of the HCD absorbers. This model also successfully characterizes the contamination effect on large scales for the highest column densities; however, on smaller scales (e.g.  $|k| > 0.4 \text{ h Mpc}^{-1}$  for small DLAs) towards the line of sight, the model fails possibly due to additional suppression in power by the most massive systems due to the effective masking of auto-correlations in the Lyman-alpha forest by their damping wings. Font-Ribera et al. (2018) found that there is much more constraining power in the 3D flux power spectrum than the 1D power spectrum for BOSS for  $|k| < 1 \text{ h Mpc}^{-1}$ , underlying the importance of accurately modelling systematics up to small scales. We therefore find that this linear Voigt model will help with precision measurements of BAO in future surveys (eBOSS/DESI) and will be essential for reconstructing the power spectrum shape beyond BAO.

## ACKNOWLEDGEMENTS

KKR, SB, HVP, and BL thank the organisers of the COSMO21 symposium in 2016, where this project was devised. KKR was

<sup>6</sup> We use this shape parameter value because the shape parameter of the BOSS model as inferred from our simulation box was considerably smaller than the BOSS best-fitting value (in order to fit the small-scale correlations)

and gave a flat response on scales larger than the size of our box (i.e. indicating no effect of damping wings in contradiction to the physical linear Voigt model; see Fig. 8).

supported by the Science and Technology Facilities Council (STFC). SB was supported by NASA through Einstein Postdoctoral Fellowship Award Number PF5-160133. HVP was partially supported by the European Research Council (ERC) under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement number 306478-CosmicDawn. AP was supported by the Royal Society. AFR was supported by an STFC Ernest Rutherford Fellowship, grant reference ST/N003853/1. BL was supported by NASA through Einstein Postdoctoral Fellowship Award Number PF6-170154. This work was partially enabled by funding from the University College London (UCL) Cosmoparticle Initiative.

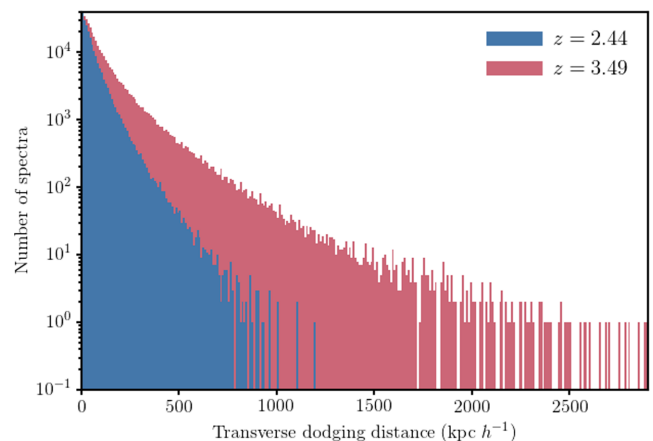
## REFERENCES

- Alcock C., Paczynski B., 1979, *Nature*, 281, 358
- Arinyo-i-Prats A., Miralda-Escudé J., Viel M., Cen R., 2015, *J. Cosmol. Astropart. Phys.*, 12, 017
- Armengaud E., Palanque-Delabrouille N., Yèche C., Marsh D. J. E., Baur J., 2017, preprint ([arXiv:e-prints](#))
- Bautista J. E. et al., 2017, *A&A*, 603, A12
- Bird S., 2017, FSFE: Fake Spectra Flux Extractor. Astrophysics Source Code Library, ascl:1710.012
- Bird S., Vogelsberger M., Haehnelt M., Sijacki D., Genel S., Torrey P., Springel V., Hernquist L., 2014, *MNRAS*, 445, 2313
- Bird S., Haehnelt M., Neeleman M., Genel S., Vogelsberger M., Hernquist L., 2015, *MNRAS*, 447, 1834
- Brooks S. P., Gelman A., 1998, *J. Comput. Graph. Stat.*, 7, 434
- Busca N. G. et al., 2013, *A&A*, 552, A96
- Dawson K. S. et al., 2013, *AJ*, 145, 10
- Dawson K. S. et al., 2016, *AJ*, 151, 44
- Delubac T. et al., 2015, *A&A*, 574, A59
- DESI Collaboration et al., 2016a, preprint ([arXiv:e-prints](#))
- DESI Collaboration et al., 2016b, preprint ([arXiv:e-prints](#))
- du Mas des Bourboux H. et al., 2017, *A&A*, 608, A130
- Eisenstein D. J. et al., 2011, *AJ*, 142, 72
- Font-Ribera A., Miralda-Escudé J., 2012, *J. Cosmol. Astropart. Phys.*, 7, 028
- Font-Ribera A., McDonald P., Miralda-Escudé J., 2012, *J. Cosmol. Astropart. Phys.*, 1, 001
- Font-Ribera A. et al., 2012, *J. Cosmol. Astropart. Phys.*, 11, 059
- Font-Ribera A. et al., 2013, *J. Cosmol. Astropart. Phys.*, 5, 018
- Font-Ribera A. et al., 2014a, *J. Cosmol. Astropart. Phys.*, 5, 027
- Font-Ribera A., McDonald P., Mostek N., Reid B. A., Seo H.-J., Slosar A., 2014b, *J. Cosmol. Astropart. Phys.*, 5, 023
- Font-Ribera A., McDonald P., Slosar A., 2018, *JCAP*, 01, 003
- Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, *PASP*, 125, 306
- Gelman A., Rubin D. B., 1992, *Statist. Sci.*, 7, 457
- Hui L., Stebbins A., Burles S., 1999, *ApJ*, 511, L5
- Iršič V., Viel M., Haehnelt M. G., Bolton J. S., Becker G. D., 2017a, preprint ([arXiv:e-prints](#))
- Iršič V. et al., 2017b, *Phys. Rev. D*, 96, 023522
- Kaiser N., 1984, *ApJ*, 284, L9
- Kaiser N., 1987, *MNRAS*, 227, 1
- Kirkby D. et al., 2013, *J. Cosmol. Astropart. Phys.*, 3, 024
- Krogager J.-K., Fynbo J. P. U., Møller P., Ledoux C., Noterdaeme P., Christensen L., Milvang-Jensen B., Sparre M., 2012, *MNRAS*, 424, L1
- Lee K.-G. et al., 2013, *AJ*, 145, 69
- McDonald P., 2003, *ApJ*, 585, 34
- McDonald P., Miralda-Escudé J., 1999, *ApJ*, 518, 24
- McDonald P., Seljak U., Cen R., Bode P., Ostriker J. P., 2005, *MNRAS*, 360, 1471
- McQuinn M., White M., 2011, *MNRAS*, 415, 2257
- Nelson D. et al., 2015, *Astron. Comput.*, 13, 12
- Palanque-Delabrouille N. et al., 2013, *A&A*, 559, A85
- Palanque-Delabrouille N. et al., 2015, *J. Cosmol. Astropart. Phys.*, 11, 011
- Pérez-Ràfols I. et al., 2018, *MNRAS*, 473, 3019
- Pontzen A., 2014, *Phys. Rev. D*, 89, 083010
- Pontzen A., Bird S., Peiris H., Verde L., 2014, *ApJ*, 792, L34
- Rogers K. K., Bird S., Peiris H. V., Pontzen A., Font-Ribera A., Leistedt B., 2018, *MNRAS*, 474, 3032
- Schaye J. et al., 2015, *MNRAS*, 446, 521
- Seljak U. et al., 2005, *Phys. Rev. D*, 71, 103515
- Slosar A. et al., 2011, *J. Cosmol. Astropart. Phys.*, 9, 001
- Slosar A. et al., 2013, *J. Cosmol. Astropart. Phys.*, 4, 026
- Vogelsberger M. et al., 2014a, *Nature*, 509, 177
- Vogelsberger M. et al., 2014b, *MNRAS*, 444, 1518
- Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, *ApJS*, 61, 249
- Yèche C., Palanque-Delabrouille N., Baur J., du Mas des Bourboux H., 2017, *JCAP*, 06, 047

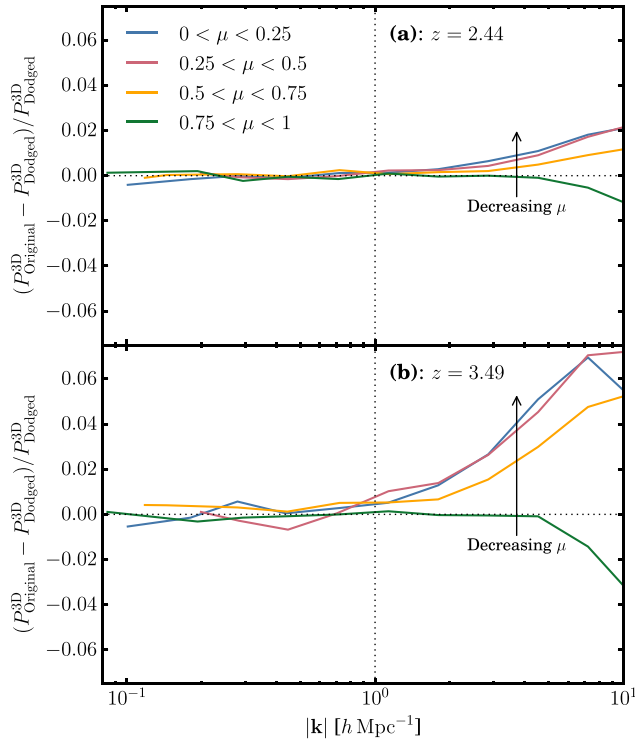
## APPENDIX A: TESTS OF HCD ABSORBER DODGING

In this appendix, we test the effect of replacing simulated spectra contaminated by HCD absorption with nearby uncontaminated spectra on our measurements of 3D flux power spectra. The measurements of these power spectra are made computationally simple by the use of FFTs, which, in turn, require a regular grid of samples. However, the transverse HCD absorber ‘dodging’ of some spectra makes this grid irregular. An error therefore arises from treating this irregular grid as the original regular grid (i.e. to ignore the transverse dodging distances) in computing the necessary FFTs.

Fig. A1 shows the distribution of the transverse dodging distances required to find replacement mock spectra uncontaminated by HCD absorbers, for the simulation boxes at the two redshifts we consider. (See Section 3.2 for more details about why and how we dodge HCD absorbers.) Replacement spectra are trialled increasingly further away from the original spectrum in steps of  $10 \text{ kpc } h^{-1}$  until an uncontaminated spectrum is found. Many of the final replacement spectra require many iterations to be found; this is exacerbated by



**Figure A1.** Histogram of the transverse comoving distances ‘dodged’ by each simulated spectrum in order to avoid HCD absorbers. The total number of spectra at each redshift  $z$  is 562 500. The number of spectra remaining un-dodged at  $z = 2.44$  and  $3.49$  is, respectively, 391 500 (69.6 per cent) and 257 063 (45.7 per cent). There is a tail of large dodging distances, much larger than the physical size of the most massive HCD absorbers because sometimes, in dodging one absorber, the proposed replacement spectrum will coincide with another absorber, somewhere else along the line of sight, requiring further dodging.



**Figure A2.** The fractional error in the estimation of the power spectrum of a GRF due to ‘dodging lines of sight’. We replicate in a GRF the exact movement of lines of sight that we carry out in our simulation box in order to dodge HCD absorbers. We then calculate the error in the estimation of the power spectrum due to ignoring the changes in position of lines of sight when calculating the necessary (fast) Fourier transforms. We note that the error remains small (sub-percent) for scales of interest in our study ( $|\mathbf{k}| < 1 h \text{ Mpc}^{-1}$ ). (a) above: we replicate the dodging in our box at  $z = 2.44$ ; (b) below: we replicate the dodging in our box at  $z = 3.49$ .

dodging one HCD absorber but then finding another HCD absorber elsewhere along the spectrum which then requires further dodging. More dodging is required at higher redshift because neutral hydrogen is more abundant.

We have tested the effect of moving some lines of sight, but then ignoring the changes in positions in the calculation of FFTs. In order to approximate our box of mock spectra, we generate a Gaussian random field (GRF) from a cosmological power spectrum (the same as input to the simulations). It is sampled at the same resolution as our mock spectra. We then estimate the power spectrum from this box, in the same way that we do with our mock spectra, forming  $P_{\text{Original}}^{\text{3D}}$ . In order to approximate our box of mock spectra after the dodging procedure, we then replicate on our GRF the exact movement of lines of sight as we carry out in our simulations. We then estimate the power spectrum from this new ‘dodged’ box, in the same way that we do with our mock spectra, i.e. ignoring the changes in positions of our replacement lines of sight. This forms  $P_{\text{Dodged}}^{\text{3D}}$ . An error is introduced in estimating the power spectrum and ignoring the transverse distances that some of the samples of the field have moved (this error could be avoided if we did not rely in the calculation of power spectra on FFTs that require evenly sampled functions). Fig. A2 shows the fractional error from this effect for the dodges we carry out at the different redshifts we consider. There is more error at higher redshift because there is more dodging (see Fig. A1), but the error remains sub-percent for scales of interest

for our study ( $|\mathbf{k}| < 1 h \text{ Mpc}^{-1}$ ). This figure should be compared to Fig. 4 (b) (the fractional effect of a mock residual contamination of HCD absorbers on the 3D Lyman-alpha forest flux power spectrum); for the scales of interest, the effect of the HCD absorbers that we wish to measure remains much larger than the error arising from the dodging. We therefore ignore only scales smaller than this cut-off in our analysis.

## APPENDIX B: DETAILS OF MODELLING AND MCMC SAMPLING

In Section 3.4, we wish to sample the joint posterior probability distribution of the parameters of our proposed model in equation (5), given our measured 3D flux power spectra. We therefore require a likelihood function for our simulated data, given the model. We use a Gaussian likelihood function and assume the covariance matrix to be diagonal (i.e. we ignore correlations between power spectrum bins). Each  $\hat{\delta}_{\text{Flux}}(\mathbf{k}_n)$  is well approximated by a Gaussian random variable and so, as explained in Section 3.3, our flux power spectrum estimates in each bin  $i$  are the sums of the squares of  $N_i$  such variables. It follows that each of the elements of our data vector is chi-squared distributed (ignoring the slightly different amplitudes of and possible correlations between the Fourier modes within each bin):<sup>7</sup>

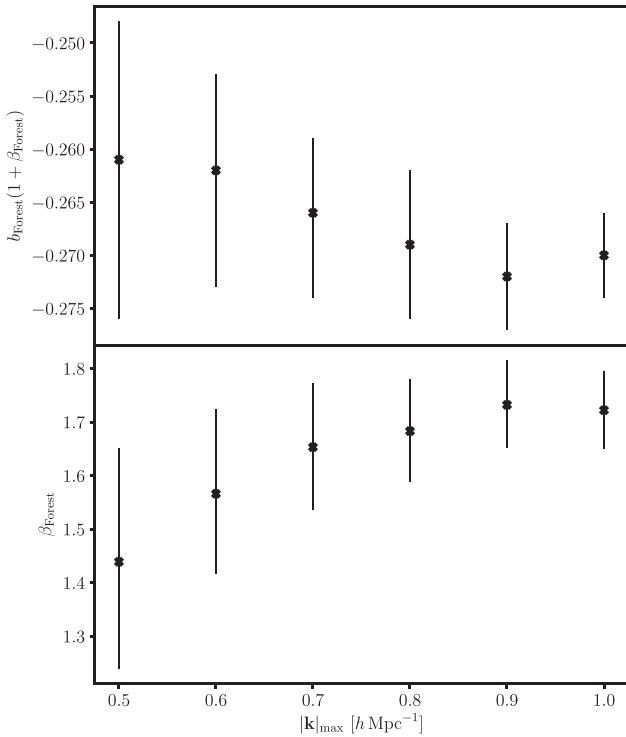
$$P_{\text{Flux},i}^{\text{3D}} \sim \frac{P_{\text{True},i}^{\text{3D}}}{N_i} \chi^2(N_i), \quad (\text{B1})$$

where  $P_{\text{True},i}^{\text{3D}}$  is the true (ensemble) value of the 3D flux power spectrum in bin  $i$ . The variance of this distribution is  $2(P_{\text{True},i}^{\text{3D}})^2/N_i$  and these form the (diagonal) elements of our covariance matrix (substituting  $P_{\text{Flux},i}^{\text{3D}}$  for  $P_{\text{True},i}^{\text{3D}}$ , which is otherwise a priori unknown). To constrain the contamination parameters while marginalizing over intrinsic Lyman-alpha forest bias parameters, we combine the likelihoods for the  $P_{\text{Contaminated}}^{\text{3D}}$  and  $P_{\text{Forest}}^{\text{3D}}$  data vectors (using equation (1) to model  $P_{\text{Forest}}^{\text{3D}}$ ); we ignore correlations between the two data vectors and simply add the log-likelihoods. This is sufficient for the level of accuracy of our study.

Our model requires evaluation of  $P_{\text{Linear}}^{\text{3D}}$  in each bin; to improve the comparison to  $P_{\text{Flux},i}^{\text{3D}}$ , we similarly evaluate  $P_{\text{Linear}}^{\text{3D}}$  at each individual mode and bin in the same way. We associate with each bin the average values of  $|\mathbf{k}|$  and  $\mu$  from the contributing modes. We use uniform prior probability distributions for the Lyman-alpha forest bias parameters  $b_{\text{Forest}}$  and  $\beta_{\text{Forest}}$  and use Gaussian priors for the HCD absorber bias parameters  $b_{\text{HCD}}$  and  $\beta_{\text{HCD}}$ , which are otherwise poorly constrained. The mean and  $1\sigma$  values of the Gaussian priors on  $b_{\text{HCD}}$  are (following the best-fitting values on the total  $b_{\text{HCD}}$  from Bautista et al. 2017) from the top to bottom, the left- to right-hand panel in Fig. 5, (a)  $-0.0005 \pm 0.0002$ ; (b)  $-0.003 \pm 0.001$ ; (c)  $-0.007 \pm 0.003$ ; (d)  $-0.016 \pm 0.006$ ; (e)  $-0.0007 \pm 0.0002$ ; (f)  $-0.005 \pm 0.002$ ; (g)  $-0.012 \pm 0.004$ ; and (h)  $-0.022 \pm 0.009$ . They are estimated by dividing the total  $b_{\text{HCD}}$  from Bautista et al. (2017) by the relative rest-frame equivalent widths of the damping wings of each absorber category; they are scaled up at higher redshift by the increased amount of HCD absorption (estimated from the fraction of contaminated spectra). Following the best-fitting values found by Bautista et al. (2017), we place a Gaussian prior on  $\beta_{\text{HCD}} = 0.7 \pm 0.2$ . Our prior distributions for these contamination parameters are almost exactly returned in their marginalized 1D

<sup>7</sup> For practical purposes, for  $N_i > 50$ , the distribution is close to a Gaussian distribution.





**Figure C1.** The maximum posterior values with the  $1\sigma$  credible intervals of the bias  $b_{\text{Forest}}$  (*above*) and redshift space distortion  $\beta_{\text{Forest}}$  (*below*) parameters of the Lyman-alpha forest, as inferred from our simulation box at  $z = 2.44$ , using different values of  $|k|_{\max}$ , the smallest scale included in our analysis. Following e.g. Slosar et al. (2011) and Bautista et al. (2017), we sample the combination  $b_{\text{Forest}}(1 + \beta_{\text{Forest}})$ , which is less correlated with  $\beta_{\text{Forest}}$ . We find that our marginalized parameter posteriors are statistically consistent, irregardless of the smallest scale at which we cut our data vector.

posterior distributions (details are given in Section 4.2), i.e. we are very insensitive to the amplitude of the effect of HCD absorbers. The shape of the scale-dependent bias is fully determined by the physics of the Voigt absorption profiles and the CDDF (see Section 2.2). We investigate the suitability of our model for the scale-dependent bias arising from the absorption profiles of HCD absorbers as a

function of column density by repeating the posterior sampling for the  $P_{\text{Contaminated}}^{\text{3D}}$  constructed for each HCD absorber category.

We sample the posterior distributions using an MCMC method, specifically EMCEE (Foreman-Mackey et al. 2013), an implementation of the affine-invariant MCMC sampler. We initialize our chains uniformly within the (non-zero) bounds of our priordistributions and test for convergence using the Gelman–Rubin statistic (Gelman & Rubin 1992; Brooks & Gelman 1998).

### APPENDIX C: TESTS OF ROBUSTNESS OF INFERENCE OF BIAS PARAMETERS OF THE LYMAN-ALPHA FOREST

Fig. C1 shows the results of testing how changing the smallest scale that we include in our analysis  $|k|_{\max}$  affects the (marginalized 1D) posterior distributions inferred for the bias parameters of the Lyman-alpha forest. The  $1\sigma$  credible intervals on the combination  $b_{\text{Forest}}(1 + \beta_{\text{Forest}})$  and  $\beta_{\text{Forest}}$  increase as  $|k|_{\max}$  decreases because the number of modes remaining on scales larger than  $|k|_{\max}$  falls off quite sharply as  $|k|_{\max}$  is reduced. Although the combination  $b_{\text{Forest}}(1 + \beta_{\text{Forest}})$  is sampled, rather than  $b_{\text{Forest}}$  alone, because it is less correlated with  $\beta_{\text{Forest}}$ , there is evidently still correlation: As  $\beta_{\text{Forest}}$  decreases with  $|k|_{\max}$ , so does also the amplitude of  $b_{\text{Forest}}(1 + \beta_{\text{Forest}})$ . None the less, the posteriors of both bias parameters are statistically consistent for all the values of  $|k|_{\max}$  that we consider, suggesting a degree of robustness in our inference on these parameters. Moreover, our conclusions on the scale-dependence of the effect of HCD absorbers on correlations in the 3D Lyman-alpha forest are insensitive to the overall amplitude, including the biases of the Lyman-alpha forest.

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.