Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **514,** 3222–3245 (2022) https://doi.org/10.1093/mnras/stac1344
Advance Access publication 2022 May 17

LyMAS reloaded: improving the predictions of the large-scale Lyman-α forest statistics from dark matter density and velocity fields

S. Peirani,[1,2]★ S. Prunet,[1,2] S. Colombi,[2] C. Pichon,[2,3] D. H. Weinberg,[4] C. Laigle,[2] G. Lavaux,[2] Y. Dubois[2] and J. Devriendt[5,6]

[1]*Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, Bd de l'Observatoire, CS 34229, F-06304 Nice Cedex 4, France*
[2]*Sorbonne Université, CNRS, UMR7095, Institut d'Astrophysique de Paris, 98 bis Boulevard Arago, F-75014 Paris, France*
[3]*IPhT, DRF-INP, UMR 3680, CEA, Orme des Merisiers Bat 774, F-91191 Gif-sur-Yvette, France*
[4]*Department of Astronomy, The Ohio State University, Columbus, OH 43210, USA*
[5]*Sub-department of Astrophysics, University of Oxford, Keble Road, Oxford OX1 3RH, UK*
[6]*Université de Lyon, Université Lyon 1, ENS de Lyon, CNRS, Centre de Recherche Astrophysique de Lyon UMR5574, F-69230 Saint-Genis-Laval, France*

Accepted 2022 May 7. Received 2022 May 6; in original form 2021 September 18

ABSTRACT

We present LyMAS2, an improved version of the 'Lyman-α Mass Association Scheme' aiming at predicting the large-scale 3D clustering statistics of the Lyman-α forest (Ly α) from moderate-resolution simulations of the dark matter (DM) distribution, with prior calibrations from high-resolution hydrodynamical simulations of smaller volumes. In this study, calibrations are derived from the HORIZON-AGN suite simulations, (100 Mpc h)$^{-3}$ comoving volume, using Wiener filtering, combining information from DM density and velocity fields (i.e. velocity dispersion, vorticity, line-of-sight 1D-divergence and 3D-divergence). All new predictions have been done at $z = 2.5$ in redshift space, while considering the spectral resolution of the SDSS-III BOSS Survey and different DM smoothing (0.3, 0.5, and 1.0 Mpc h^{-1} comoving). We have tried different combinations of DM fields and found that LyMAS2, applied to the HORIZON-NOAGN DM fields, significantly improves the predictions of the Ly α 3D clustering statistics, especially when the DM overdensity is associated with the velocity dispersion or the vorticity fields. Compared to the hydrodynamical simulation trends, the two-point correlation functions of pseudo-spectra generated with LyMAS2 can be recovered with relative differences of ∼5 per cent even for high angles, the flux 1D power spectrum (along the light of sight) with ∼2 per cent and the flux 1D probability distribution function exactly. Finally, we have produced several large mock BOSS spectra (1.0 and 1.5 Gpc h^{-1}) expected to lead to much more reliable and accurate theoretical predictions.

Key words: methods: numerical – dark matter.

1 INTRODUCTION

Distant quasars emit light that crosses a large part of the Universe before being observed with instruments on Earth. In particular, the spectrum of each quasar presents fluctuating absorption that corresponds to the Lyman-α forest (Ly α; Lynds 1971; Sargent et al. 1980). The study of the Ly α forest has become a major focus of modern cosmology, as it is supposed to trace the neutral hydrogen density that fills most of the Universe in a way that approximately corresponds to the underlying dark matter (DM) density (Croft et al. 1999; Peeples et al. 2010). Since a single background source only provides 1D information along the corresponding line of sight (LOS, or 'skewer'), characterizing the 3D density of the high-redshift Universe with the Ly α forest requires large samples of quasar spectra. Successful surveys such as the (extended) Baryon Oscillation Spectroscopic Survey (BOSS/eBOSS; Dawson et al. 2013, 2016), of the Sloan Digital Sky Survey (SDSS-III and SDSS-IV; Blanton et al. 2017), Eisenstein et al. (2011) have measured the Ly α forest spectra of 160 000 quasars at redshifts $2.2 < z < 3$. Thanks to this large sample, the study of the Ly α forest has proved to be a complementary probe to low-redshift galaxy surveys. For instance, the large sample of quasar spectra have permitted accurate measurements of 3D flux autocorrelation functions (Slosar et al. 2011) as well as the cross-correlation between the Ly α Forest and specific tracers, namely damped-Ly α systems (DLAs) and quasars (Font-Ribera et al. 2012, 2013, 2014). Such 3D measurements also enable measurements of the distance–redshift relation and the Hubble expansion via baryon accoustic oscillations (BAO; Busca et al. 2013; Slosar et al. 2013; Delubac et al. 2015; Bautista et al. 2017; du Mas des Bourboux et al. 2020). Moreover, BOSS spectra also permit accurate measurements of the LOS power spectrum (Palanque-Delabrouille et al. 2013) and flux probability distribution function (PDF; Lee et al. 2015). In the near future, the *Dark Energy Spectroscopic Instrument* (DESI; DESI Collaboration et al. 2016), the *William Herschel Telescope Enhanced Area Velocity Explorer* (WEAVE-QSO; Dalton et al. 2016, 2020; Pieri et al. 2016), and the *Subaru Prime Focus Spectrograph* (PFS; Takada et al. 2014) will go well beyond the present surveys and will open new perspectives on the high redshift intergalactic medium probed by the Ly α forest.

In parallel to the development of these large quasar surveys, theoretical modelling needs to reach the high level of complexity and accuracy to interpret the observational data. Nowadays, hydrodynamical cosmological simulations represent an ideal tool as they manage

★ E-mail: sebastien.peirani@oca.eu

to model the intergalactic medium with a high degree of realism with appropriate resolution (e.g. Dubois et al. 2014; Vogelsberger et al. 2014; Schaye et al. 2015; Bolton et al. 2017). However, to properly model the 3D correlations of the Ly α forest, one needs to resolve the pressure-support scale (Jeans scale) of the diffuse intergalactic medium (IGM), which is typically $\sim$0.25 Mpc h^{-1} comoving for matter overdensity of $\sim$10 (Peeples et al. 2010), while considering $\sim$Gpc^3 simulation volumes to exploit the statistical precision achieved by the different observational surveys while avoiding box size effects. Combining such resolution and simulation volume is currently not feasible mainly because of computational limits. To tackle such an issue, several methods exist in the literature. One of the most popular is to use the so-called 'Fluctuating Gunn–Peterson Approximation' (FGPA; Weinberg et al. 1997; Croft et al. 1998) that links the Ly α optical depth to the local DM density. This approach is relatively straightforward as it assumes a deterministic relation and only information on the density field (extracted from *N*-body simulations or lognormal density fields) is required. However, the FGPA approach is expected to be accurate enough only on very large scales, e.g. those of the BAO features ($\sim$ 100 Mpc h^{-1}; e.g. fig. 5 of Sinigaglia et al. 2022). But 3D Ly α forest surveys also enable precise measurements of flux correlations at much smaller scales where the FGPA might not be adequate.

Another approach is to apply relevant calibrations, derived first from small hydrodynamical simulations, to large-scale DM distributions extracted from pure DM simulations, which are much cheaper to perform. In particular, Peirani et al. (2014, hereafter P14) have developed the Lyman-α Mass Association Scheme (LyMAS) that follows such a philosophy. The main idea is that flux correlations on small and large scales are mainly driven by the correlations of the DM density field. More specifically, the flux statistics can be estimated by combining the DM density field with the conditional probability distribution $P(F|\rho)$ of the transmitted flux F on the DM density contrast ρ. In its most sophisticated form, LyMAS creates ensemble of coherent pseudo-spectra at the BOSS resolution using the Gaussianized percentile distribution of the conditional flux, while re-scaling the LOS power spectrum and PDF at the last step. One of the main results of LyMAS is to improve the predictions of flux 3D correlations especially with respect to deterministic mapping (e.g. FGPA) that tends to significantly overestimate them especially when the DM density is smoothed at scale greater than 0.3 Mpc h^{-1}. Similarly, Sorini et al. (2016) have developed '*Iterative Matched Statistics*' (IMS) in which the PDF and the power spectrum of the real-space Ly α flux are derived from small hydrodynamical simulations. Then, these two statistics are 1D (1D-IMS) or 3D (3D-IMS) iteratively mapped on to a pseudo-flux field of an *N*-body simulation from which the matter density is first Gaussian smoothed. In 3D-IMS, smoothing is followed by matching the 3D power spectrum and PDF of the flux in real space to the reference hydrodynamic simulation. With 1D-IMS, the 1D power spectrum and PDF of the flux are additionally matched. Both methods have proved to be again more accurate than the FGPA approach (which strongly relies on the DM smoothing scale) when reproducing LOS observables, such as the PDF and power spectrum as well as the 3D flux power spectrum (5–20 per cent). Finally, machine-learning-based methods start to be considered and lead to promising results (Chopitan, Lavaux & Peirani, in preparation; Harrington et al. 2022; Sinigaglia et al. 2022).

Although the LyMAS full scheme is able to model the BOSS 3D clustering quite accurately and has been already used in different analysis related to the quasar-Ly α forest cross-correlation (Lochhaas et al. 2016), the three-point correlation functions (Tie et al. 2019) and the correlations between the Ly α transmitted flux and the mass overdensity (Cai et al. 2016), we aim at investigating whether other sets of calibrations could still improve the theoretical predictions. To this regard, we consider a new approach based on Wiener Filtering, which has been used for 3D map reconstruction from an ensemble of 1D LOS (e.g. Pichon et al. 2001; Caucci et al. 2008; Ozbek, Croft & Khandai 2016; Lee et al. 2018; Japelj et al. 2019; Ravoux et al. 2020). The underlying philosophy in LyMAS2 remains unchanged. We still find that the flux correlations are driven mainly by the correlations of the DM density field, but with potential refinements from the correlations of the DM velocity field.

This paper is organized as follows. In Section 3, we describe the Wiener equations as multivariate normal conditional probabilities, and we explain their application to hydrodynamical simulations. Section 4 briefly describes how we extract the flux and all relevant DM fields from the HORIZON-NOAGN simulation. We also present the potential correlations that arise between these different fields. Then, in Section 5, we present the statistics in the LOS power spectrum, the PDF and the two-point correlation function of pseudo-spectra produced when LyMAS2 is applied to different associations of DM fields of HORIZON-NOAGN. Such trends are compared to the flux statistics from the hydrodynamical simulation ('hydro flux'). In Section 6, we apply LyMAS2 to *N*-body simulations of 1.0 and 1.5 Gpc h^{-1} comoving volumes. We summarize our results and conclusions in Section 7. We also add three appendices. In Appendix A, we compute the mean two point correlations functions from five different hydrodynamical simulations of lower resolution to check the robustness of the results presented in Section 5. Appendix B presents the performance of specific deterministic samplings. Appendix C provides details on how estimates of density and velocities are performed on the DM particle distribution, relying on adaptive softening.

2 LYMAS VERSUS LYMAS2

We briefly describe the fundamental differences between the first version of LyMAS, detailed in P14 and the new scheme, LyMAS2, presented in this work. The two versions basically share the same philosophy: specific calibrations are first derived from hydrodynamical simulations of small volume and then applied to large DM simulations, assuming that the correlations of the Ly α at small and large scales are mainly driven by the correlations of the underlying DM density and (eventually) velocity fields. The main differences essentially lie in (1) the derivation and characterization of the cross-correlation between the different fields (namely the hydro spectra and the DM fields) and (2) the way we apply such calibrations to the DM distributions.

More specifically, in the first version a hydrodynamical simulation was used to calibrate the conditional probability distribution $P(F|\rho)$ to have a transmitted flux value F, given the value of the DM density contrast ρ at the same location. In its simplest form, LyMAS randomly and independently draws transmitted flux values according to $P(F|\rho)$ and the value of ρ at each pixel of a regular grid used to sample the DM overdensity field. Although the 3D clustering statistics of the pseudo-spectra generated by this approach is quite close to that of the hydro flux, the main drawback of this procedure is to create very noisy spectra as any coherence along each LOS is lost. To solve this issue and make the pseudo-spectra more realistic, the most sophisticated form of LyMAS uses the fact that neighbouring pixels along a given LOS are supposed to have close probability distribution $P(F|\rho)$. Hence, one can introduce a coherence by defining percentile spectra, i.e. the fractional position

of the flux value in the cumulative distribution of $P(F|\rho)$. Then these percentile spectra derived from every LOS of the grid can be Gaussianized and one can derive a characteristic power spectrum from these 1D Gaussian fields. Thus, from this new input parameter, LyMAS generates first a 1D Gaussian field, de-Gaussiannizes it to get a realization of a percentile spectrum, and finally derives a coherent spectrum using the different values of the DM density contrast ρ and the percentile value in $P(F|\rho)$ along the considered LOS. Here again, the predictions of the 3D clustering from such coherent spectra is proved be very accurate.

LyMAS2 does not derive and consider $P(F|\rho)$ as well as percentile spectra. Instead, as we explain in detail in Section 3, LyMAS2 makes good use of Wiener Filtering to characterize the correlations between the transmitted flux and the DM density contrast. The statistics are directly made LOS by LOS, which naturally introduces a coherence in the pseudo-spectra. Furthermore, this approach has the advantage to naturally take into account not only the DM density field (like in LyMAS) but other fields such as specific DM velocity fields (e.g. velocity dispersion, vorticity, divergence) that potentially bring new information to improve the predictions.

In the very last step, both LyMAS and LyMAS2 end similarly by rescaling the flux LOS power spectrum and PDF. These transformations are useful to slightly correct the LOS 1D power spectrum and PDF of pseudo-spectra to make them identical or quasi identical to those of the hydro spectra. This step, however, does not significantly modify the 3D clustering statistics.

In the beginning of Section 5, we summarize all the steps of LyMAS and LyMAS2 to create a pseudo-spectrum.

3 WIENER EQUATIONS

3.1 Multivariate conditional probabilities

Let us assume a complex Gaussian random (vector) variable $\boldsymbol{x}$ that can be separated into two sub-vectors $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, whose mean and covariance can be written as

$$\mu = (\mu_1, \mu_2),$$
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^H & \Sigma_{22} \end{pmatrix},$$

where the H superscript denotes the Hermitian conjugate. By using formulas for block inverses, it is possible to derive from the joint distribution of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ the formula for the conditional distribution of $\boldsymbol{x}_1$, given $\boldsymbol{x}_2$. As expected, it is a Gaussian multivariate distribution, of mean and covariance:

$$\bar{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)\,, \tag{1}$$

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^H\,. \tag{2}$$

Now, consider the joint Gaussian distribution of the (complex) spatial Fourier modes $(f_k, \delta_k, \theta_k, \omega_k)$, where, by definition for field a, we have $a_k = \int a(x)\mathrm{e}^{-ik\cdot x}\mathrm{d}^3x$. We assume they are centred fields (zero mean), and of covariance (we drop the subscript k for visibility):

$$\Sigma = \left(\begin{array}{cccc} P_{ff} & P_{f\delta} & P_{f\theta} & P_{f\omega} \\ \hline P_{f\delta}^* & P_{\delta\delta} & P_{\delta\theta} & P_{\delta\omega} \\ P_{f\theta}^* & P_{\delta\theta}^* & P_{\theta\theta} & P_{\theta\omega} \\ P_{f\omega}^* & P_{\delta\omega}^* & P_{\theta\omega}^* & P_{\omega\omega} \end{array}\right) = \left(\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{12}^H & \Sigma_{22} \end{array}\right),$$

where, by definition, $P_{ab}(k) \equiv \langle a_k^* b_k\rangle$ is the cross-spectrum of fields a and b at wavenumber k, and we have partitioned the fields according to $\boldsymbol{x}_1 \equiv f_k$ and $\boldsymbol{x}_2 \equiv (\delta_k, \theta_k, \omega_k)^T$. Applying the equations (1) and (2), we obtain the conditional mean and variance of the field f_k, given $(\delta_k, \theta_k, \omega_k)$:

$$\bar{f}_k = \left(P_{f\delta}, P_{f\theta}, P_{f\omega}\right)\cdot\boldsymbol{\Sigma}_{22}^{-1}\cdot\begin{pmatrix}\delta_k\\ \theta_k\\ \omega_k\end{pmatrix}, \tag{3}$$

$$\bar{\boldsymbol{\Sigma}} = P_{ff} - \left(P_{f\delta}, P_{f\theta}, P_{f\omega}\right)\cdot\boldsymbol{\Sigma}_{22}^{-1}\cdot\begin{pmatrix}P_{f\delta}^*\\ P_{f\theta}^*\\ P_{f\omega}^*\end{pmatrix}. \tag{4}$$

Computing the inverse $\boldsymbol{\Sigma}_{22}^{-1}$ using the cofactor matrix formula, we obtain

$$\Sigma_{22}^{-1} = \frac{1}{|\Sigma_{22}|}\begin{pmatrix}A_{11} & A_{12} & A_{13}\\ A_{21} & A_{22} & A_{23}\\ A_{31} & A_{32} & A_{33}\end{pmatrix},$$

with

$$\begin{aligned}
A_{11} &= P_{\theta\theta}P_{\omega\omega} - P_{\theta\omega}P_{\theta\omega}^*\,, & A_{12} &= -P_{\delta\theta}^*P_{\omega\omega} + P_{\delta\omega}^*P_{\theta\omega}\,,\\
A_{13} &= P_{\delta\theta}^*P_{\theta\omega}^* - P_{\delta\omega}^*P_{\theta\theta}\,, & A_{21} &= -P_{\delta\theta}P_{\omega\omega} + P_{\delta\omega}P_{\theta\omega}^*\,,\\
A_{22} &= P_{\delta\delta}P_{\omega\omega} - P_{\delta\omega}P_{\delta\omega}^*\,, & A_{23} &= -P_{\delta\delta}P_{\theta\omega}^* + P_{\delta\theta}P_{\delta\omega}^*\,,\\
A_{31} &= P_{\delta\theta}P_{\theta\omega} - P_{\theta\theta}P_{\delta\omega}\,, & A_{32} &= -P_{\delta\delta}P_{\theta\omega} + P_{\delta\theta}^*P_{\delta\omega}\,,\\
A_{33} &= P_{\delta\delta}P_{\theta\theta} - P_{\delta\theta}P_{\delta\theta}^*\,,
\end{aligned}$$

and

$$\begin{aligned}
|\Sigma_{22}| = {} & P_{\delta\delta}(P_{\theta\theta}P_{\omega\omega} - P_{\theta\omega}P_{\theta\omega}^*) + (P_{\delta\theta}P_{\theta\omega}P_{\delta\omega}^* + \text{c.c.})\\
& - P_{\theta\theta}P_{\delta\omega}P_{\delta\omega}^* - P_{\omega\omega}P_{\delta\theta}P_{\delta\theta}^*\,,
\end{aligned}$$

where 'c.c' denotes the conjugate complex.

Note that we limit our study to a maximum of three different input fields (i.e. δ, θ, and ω) to construct the field f. But obviously, this formalism can be extended to a higher number of fields, leading to more and more complex analytical solution. However, we will see that the statistical trends derived when considering two and three input fields are quite similar (when judiciously chosen), suggesting that adding more than two fields will not noticeably improve the results anymore.

3.2 Application to hydro simulations

Let us call F, ρ, v_1, and v_2, respectively, the local Ly α transmitted flux, the local DM density, velocity divergence, and vorticity amplitude, extracted from a given hydrodynamical simulation. Let us call now G the cumulative distribution of a standard normal $\mathcal{N}(0;1)$ distribution ($G[x] = \int^x \exp(-u^2/2)/\sqrt{2\pi}\mathrm{d}u$), and

$$G_\rho = \int_{-\infty}^{\rho}\mathrm{PDF}(\rho')\mathrm{d}\rho', \tag{5}$$

$$G_{v_i} = \int_{-\infty}^{v_i}\mathrm{PDF}(v')dv', \tag{6}$$

$$G_\mathrm{F} = \int_0^F \mathrm{PDF}(F')\mathrm{d}F'\,, \tag{7}$$

the cumulative distributions of the measured DM density and velocity fields (in the hydrodynamical simulation), and G_f the cumulative distribution of the measured flux, F. We can then define new fields, namely

$$f = G^{-1}(G_\mathrm{F}(F), \delta = G^{-1}(G_\rho(\rho)), \tag{8}$$

$$\theta = G^{-1}(G_{v_1}(v_1)), \omega = G^{-1}(G_{v_2}(v_2))\,. \tag{9}$$

which should be normally distributed by construction (or 'gaussianized'). Let us compute these different fields from the hydro

simulations and extract from them the relevant cross-spectra, using Fourier space:

$$P_{ff} = \langle f_k^* f_k \rangle, \quad P_{\delta\delta} = \langle \delta_k^* \delta_k \rangle, \quad P_{\theta\theta} = \langle \theta_k^* \theta_k \rangle,$$
$$P_{\omega\omega} = \langle \omega_k^* \omega_k \rangle, \quad P_{f\delta} = \langle f_k^* \delta_k \rangle, \quad P_{f\theta} = \langle f_k^* \theta_k \rangle,$$
$$P_{\delta\theta} = \langle \delta_k^* \theta_k \rangle, \quad P_{\theta\omega} = \langle \theta_k^* \omega_k \rangle, \text{ etc.}$$

which are going to depend typically on a transverse and a longitudinal separation radius. If we assume that the fields f, δ, θ, and ω are Gaussian random fields (GRFs, not just its one point statistics is now required to be normal), then for a given measured set of ρ, v_1, and v_2 (correspondingly a set of δ_k, θ_k, and ω_k) ,say along a set of LOS, one can estimate the most likely field $\bar{f}$ following equations (3):

$$\bar{f}_k = T_1 \cdot \delta_k + T_2 \cdot \theta_k + T_3 \cdot \omega_k \,, \tag{10}$$

where T_1, T_2, and T_3 are functions of the cross-spectrum $P_{ab}(k)$. This approach can be done along a given LOS or in volume. However, in this study we will only analyse LOS individually and independently, ignoring transverse correlations between different LOS for now. This allows us to use 1D fast Fourier transforms (FFT), and assuming stationarity along the LOS, the multiplication is simply done frequency by frequency since in Fourier space the covariance sub-blocks P_{ab} are diagonal. To illustrate, if only one field is considered (i.e. the DM overdensity field in our study), we simply obtain

$$\bar{f}_k = P_{f\delta} \cdot P_{\delta\delta}^{-1} \cdot \delta_k \,. \tag{11}$$

If we add additional information from a specific velocity field (e.g. velocity dispersion), the expression of $\bar{f}$ becomes

$$\bar{f}_k = \frac{P_{f\delta}(P_{\theta\theta} - P_{\delta\theta})}{P_{\delta\delta}P_{\theta\theta} - P_{\delta\theta}P_{\delta\theta}^*} \cdot \delta_k + \frac{P_{f\theta}(P_{\delta\delta} - P_{\delta\theta}^*)}{P_{\delta\delta}P_{\theta\theta} - P_{\delta\theta}P_{\delta\theta}^*} \cdot \theta_k \,. \tag{12}$$

Adding a second velocity field leads to a more complex expression of $\bar{f}_k$.

We can then draw samples as $\tilde{f}_k \equiv \bar{f}_k + \Delta f_k$, where Δf_k obeys a GRF of mean zero and variance as equation (4):

$$\Delta f_k \sim \mathcal{G}(0, P_{ff} - P_{f\delta} \cdot P_{\delta\delta}^{-1} \cdot P_{f\delta}^*) \,,$$

when only the DM density is considered and

$$\Delta f_k \sim \mathcal{G}\left(0, P_{ff} - \begin{pmatrix} P_{f\delta} \\ P_{f\theta} \end{pmatrix}^T \cdot \begin{bmatrix} P_{\delta\delta} & P_{\delta\theta} \\ P_{\delta\theta}^* & P_{\theta\theta} \end{bmatrix}^{-1} \cdot \begin{pmatrix} P_{f\delta}^* \\ P_{f\theta}^* \end{pmatrix}\right),$$

in the case of two DM fields. After computing the inverse Fourier transform of $\tilde{f}_k$ to obtain $\tilde{f}$, the corresponding flux obeys $\tilde{F} = G_f[G^{-1}(\tilde{f})]$. By construction the one point statistics of $\tilde{f}$ will be random normal, so that the one-point PDF of its de-Gaussianized transform will be that of the original field. The power spectrum $P_{\tilde{f}\tilde{f}}$ of $\tilde{f}$ will be the same as that of f. Indeed, let us consider the case with only one input file for simplicity. We have

$$P_{\tilde{f}_k\tilde{f}_k} \equiv \langle |\bar{f}_k + \Delta f_k|^2 \rangle = \left\langle \left| \frac{1}{P_{\delta\delta}} P_{f\delta}\delta_k \right|^2 + |\Delta f_k|^2 \right\rangle,$$

because the expectation and the fluctuations are uncorrelated. Therefore

$$P_{\tilde{f}_k\tilde{f}_k} = \frac{P_{f\delta}^2}{P_{\delta\delta}^2} \langle |\delta_k|^2 \rangle + P_{ff} - \frac{P_{f\delta}^2}{P_{\delta\delta}} = P_{ff} \,,$$

since $\langle |\delta_k|^2 \rangle \equiv P_{\delta\delta}$.

Recall that all equations above are valid independently for each Fourier mode k, and for each mode, all P_{ab} terms are scalars for a given pair of fields (a, b).

4 FLUX AND DM FIELDS

Throughout the present analysis, we have used the HORIZON-NOAGN simulation (Peirani et al. 2017) to characterize any relevant cross-correlations between the transmitted flux and the different DM fields. HORIZON-NOAGN is a hydrodynamical simulation of 100 h^{-1}Mpc comoving boxside run with the RAMSES code (Teyssier 2002). It evolves 1024^3 DM particles with a mass resolution of 8.27×10^7 M$_\odot$, while the initially uniform grid is refined in an adaptive way down to $\Delta x = 1$ proper kpc at all times. The simulation adopts a standard ΛCDM cosmology compatible with *WMAP*-7 results (Komatsu et al. 2011), namely a total matter density $\Omega_{\rm m} = 0.272$, a dark energy density $\Omega_\Lambda = 0.728$, an amplitude of the matter power spectrum $\sigma_8 = 0.81$, a baryon density $\Omega_{\rm b} = 0.045$, a Hubble constant $H_0 = 70.4\,{\rm km\,s^{-1}\,Mpc^{-1}}$, and $n_{\rm s} = 0.967$. HORIZON-NOAGN is the twin simulation of HORIZON-AGN (Dubois et al. 2014). It contains all relevant physical processes such as metal-dependent cooling, photoionization, and heating from a UV background, supernova feedback, and metal enrichment, but does not include black hole growth and therefore AGN feedback.

The choice of using HORIZON-NOAGN instead of HORIZON-AGN was mainly motivated by the fact that we have performed five additional but slightly lower resolution hydrodynamical simulations to estimate the accuracy and robustness of the LyMAS2 predictions. Thus, turning out the AGN feedback processes in the simulations has permitted us to limit the computational time. These results are presented in Appendix A. Furthermore, we have tuned in this study the UV background intensity in the process of generating the 'noAGN' flux grid (see below) in order to get the same mean transmitted Lyα forest flux $\overline{F}$ derived from HORIZON-AGN. By doing this, the flux statistical predictions from the two simulations in the 3D Lyα clustering tend to be almost the same. This has been already noticed in Lochhaas et al. (2016) when studying the cross-correlations between DM haloes and transmitted flux in the Lyα forest. Note, however, that AGN feedback is expected to have non-negligible effect on the Lyα 3D clustering such as, for instance, the 1D power spectrum of the Lyα forest (e.g. Viel, Schaye & Booth 2013; Chabanier et al. 2020).

In the following, we describe briefly how we derived the hydro spectra field and the different DM density and velocity fields from the HORIZON-NOAGN simulation. Similarly to P14, we analyse the simulation outputs at redshift 2.5. Each field is sampled in a regular grid of 1024^3 pixels and the size of a single pixel is therefore $\sim$0.1 comoving Mpc h^{-1} or 0.04 physical Mpc.

4.1 Transmitted flux

From the HORIZON-NOAGN, we follow the method to generate the hydro spectra that is fully described in P14. The optical depth of Lyα absorption is calculated based on the neutral hydrogen density along each LOS. Basically, the opacity at observer-frame frequency $\nu_{\rm obs}$ is $\tau(\nu_{\rm obs}) = \sum_{\rm cells} n_{\rm HI}\sigma(\nu_{\rm obs}){\rm d}l$, where the sum extends over all cells traversed by the LOS, $n_{\rm HI}$ is the numerical density of neutral H atoms in each cell, $\sigma(\nu_{\rm obs})$ is the cross-section of Hydrogen to Lyα photons, and dl is the physical size of the cell. Then each spectrum is smoothed with a 1D Gaussian of dispersion 0.696 h^{-1} Mpc, equivalent to BOSS spectral resolution at $z \approx 2.5$. The optical depth along each spectrum is converted to Lyα forest flux by $F = e^{-\tau}$. Following common practice in Lyα forest modelling, the UV background intensity is chosen to give a mean transmitted Lyα forest flux $\overline{F} = \langle e^{-\tau} \rangle = 0.795$, matching the metal-corrected $z = 2.5$ value measured from high-resolution spectra by Faucher-Giguère et al. (2008).

4.2 Density, velocity, and mean square velocity

DM skewers that correspond to the 'hydro' spectra are also extracted from the hydrodynamical simulation. We use the same three-step algorithm introduced in P14 to derive both the overdensity, the velocity field and the velocity dispersion fields:

(i) adaptive interpolation of the DM particle distribution on a regular grid (Colombi, Chodorowski & Teyssier 2007), as detailed in Appendix C;
(ii) smoothing with a Gaussian window in Fourier space;
(iii) extraction of the skewers from a grid of LOS aligned along the z-axis.

In step (ii), DM field is 3D smoothed using different choices of smoothing scales. In P14, we found that a smoothing scale of 0.3 Mpc h^{-1} has proved to be optimal leading to the most accurate predictions. However, we prefer a value of 0.5 Mpc h^{-1} in this study since the predictions are very similar to those obtained with 0.3 Mpc h^{-1} (see Appendix A). Furthermore, we anticipate with the fact that it is computationally much easier to smooth a DM field to 0.5 than 0.3 Mpc h^{-1} when considering large volumes namely with boxsides at least greater or equal to 1 Gpc h^{-1}.

4.3 DM vorticity

The velocity field is projected (using Cloud-in-Cell interpolations) on a regular grid of resolution 1024 and smoothed over 0.5 Mpc h^{-1} with a Gaussian filter. The vorticity Ω is then computed as being the curl of the velocity field using FFT. Slightly smoothing the input velocity field allows us to avoid Gibbs artefacts.

4.4 DM velocity divergence

We have considered both the 3D velocity divergence and the 1D velocity divergence along the LOS direction.

For the 3D case, we employed two different methods to see whether this could affect our results and trends. The first one is based on a centred finite-difference approximation, namely the divergence $\nabla^i_{\rm 3D}$ at a pixel i is given by

$$\nabla^i_{\rm 3D} \simeq \frac{V_x^{i+1}-V_x^{i-1}}{2h} + \frac{V_y^{i+1}-V_y^{i-1}}{2h} + \frac{V_z^{i+1}-V_z^{i-1}}{2h}, \quad (13)$$

where V_x^i, V_y^i, and V_z^i are the velocity components at pixel i and h the size of a pixel (i.e. 100/1024 Mpc h^{-1} here). The second method uses the exact expression of the divergence in Fourier space. However, we found that the two methods lead to very similar results so we will only show results from the Fourier space method for the 3D case.

For the 1D case, we simply use the finite-difference approach and the divergence $\nabla^i_{\rm 1D}$ at a pixel i becomes

$$\nabla^i_{\rm 1D} \simeq \frac{V_z^{i+1} - V_z^{i-1}}{2h}, \quad (14)$$

since we define the z-axis as the direction of the LOS.

We summarize in Table 1 the different DM fields used in this study. It is worth mentioning that we changed some of the notations that can be found in P14. We first replaced the definition of the smoothed flux at the BOSS resolution $F_{\rm s}$ in P14 to F here. We also changed the definition of the 3D smoothed DM overdensity, $\rho = 1 + \delta$ instead of $\Delta_{\rm s} = 1 + \delta_{\rm s}$ in P14.

In Fig. 1, we show a slice of the hydro flux (smoothed at BOSS resolution) and the corresponding DM density and velocity fields (smoothed at 0.5 Mpc h^{-1}) extracted from the HORIZON-NOAGN simulation at redshift 2.5. As expected, clear correlations are noticeable between the transmitted flux and the different DM fields. This trend can also be seen when studying the evolution of each field along the same LOS, and a typical example is given in Fig. 2. We note that high absorptions in the flux correspond to high-density regions or high values in the vorticity or the velocity dispersion. But

Table 1. Summary of fields and corresponding notation used in the text.

Hydro spectra	
Flux (smoothed at BOSS res.)	F
Optical depth	$\tau = -\ln F$
Dark matter fields	
Smoothed density	$\rho_{\rm s}$
Overdensity	$\rho = 1 + \delta = \rho_{\rm s}/\langle\rho_{\rm s}\rangle$
Vorticity	Ω
Velocity dispersion	σ
1D vel. divergence	$\nabla_{\rm 1D}$
3D vel. divergence	$\nabla_{\rm 3D}$

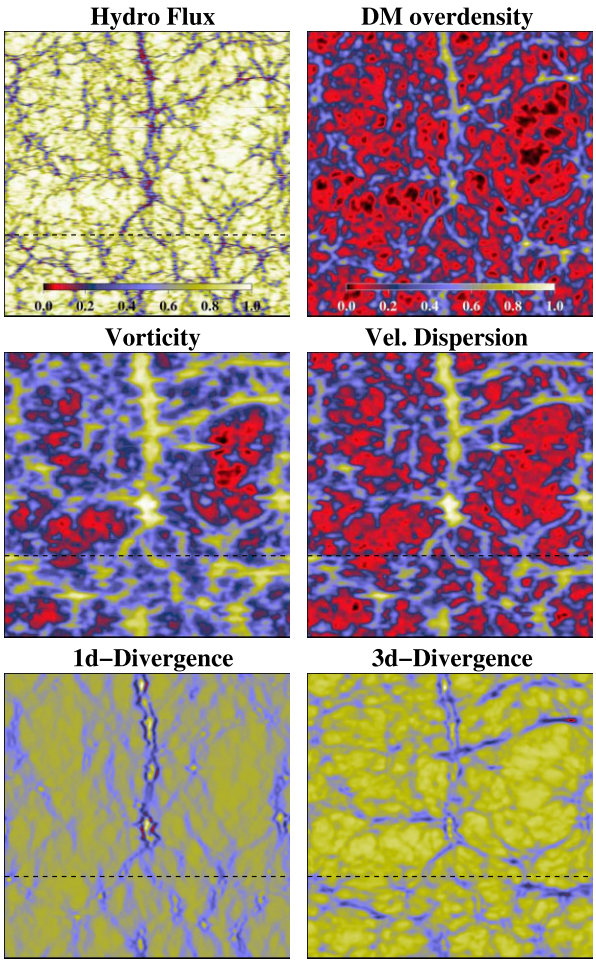

Figure 1. Slices through the flux (1D smoothed at the BOSS resolution) as well as corresponding DM overdensity and velocity fields (3D smoothed at 0.5 Mpc h^{-1}) in redshift space (horizontal direction). Each field has been extracted from the HORIZON-NOAGN simulation at $z = 2.5$ and has been normalized to help the visual comparison. The dashed lines correspond to the same LOS (see Fig. 2).

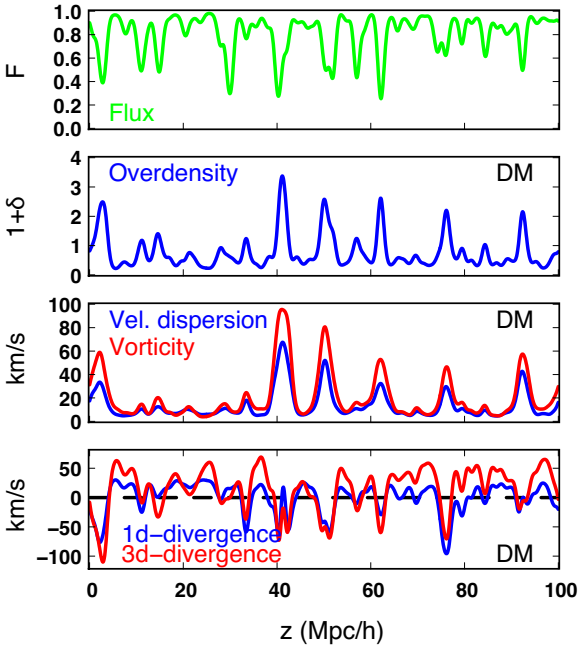

Figure 2. An example showing clear correlations (or anticorrelations) between the evolution of the hydro flux (green line) and DM density and velocity fields (blue and red lines) along the same skewer extracted from the HORIZON-NOAGN at $z = 2.5$. These skewers are extracted from the slices studied in Fig. 1.

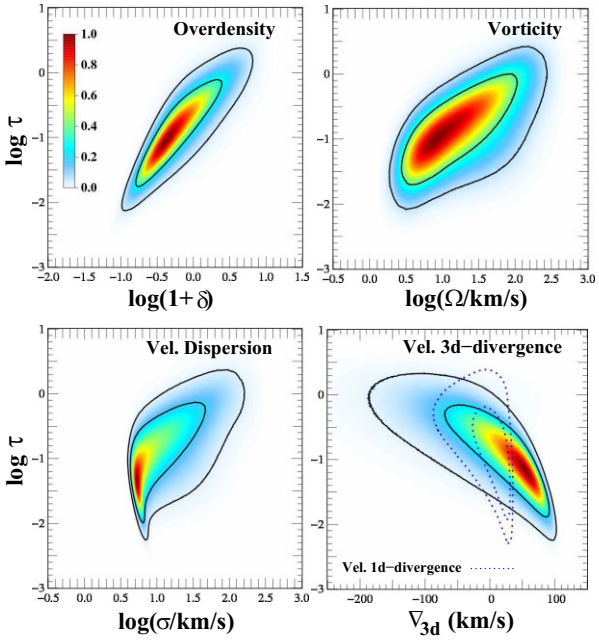

Figure 3. Correlations between the optical depth $\tau = -\ln F$ in the hydro spectra (smoothed at the BOSS resolution) and DM quantities smoothed at 0.5 Mpc h^{-1} namely the overdensity $(1 + \delta)$, the vorticity (Ω), the velocity dispersion (σ), the 3D-velocity divergence, and the 1D-velocity divergence (along the LOS) at $z = 2.5$. Colours show the density of pixel using normalized values and contour line mark areas enclosing 68.27 and 95.45 per cent. Hydro spectra and DM fields have been computed in redshift space and extracted from the HORIZON-NOAGN simulation.

the relative amplitudes of peaks in the density contrast may differ from those of the the velocity dispersion/vorticity. Indeed, the density contrast and the velocity dispersion/vorticity do not necessary put the emphasis of the same structures (e.g. walls, filaments) as suggested by Fig. 1 or, for instance, by fig. 2 of Buehlmann & Hahn (2019). In contrast, these high absorptions rather coincide with high negative values in the 3D or 1D velocity divergence. This is consistent since high-density regions are associated with DM haloes in which matter tends to sink toward the centre of the objects. Note also that the variations of the modulus of the vorticity and velocity dispersion are very similar.

4.5 Field cross correlations

In order to characterize the correlations that emerge from Figs 1 and 2, we first plot in Fig. 3 some relevant scatter plots between the optical depth τ and the DM overdensity and velocity fields. The correlations between the optical depth in the hydro spectra and the smoothed DM overdensity $(1 + \delta)$ is quite similar to the trend found in P14 using the 50 h^{-1} Mpc 'Horizon MareNostrum' simulation. In Fig. 4, we additionally show the correlations between the different DM fields. As noticed in Fig. 2, the velocity dispersion and vorticity field are highly correlated. We do not show the correlations using the 1D velocity divergence as there are quite similar with trends found using the 3D velocity divergence.

All these plots suggest that there are more or less pronounced correlations between the different input DM fields. It is however tricky to anticipate which combinations of fields through the LyMAS2 scheme would lead to the most accurate theoretical predictions. As specified in Section 3, we consider combinations with up to three different DM fields, which offers 85 different possibilities (5, 20, and 60, respectively, for one, two, and three input fields). However, as the main philosophy of LyMAS is to trace the Ly α flux from the underlying DM distribution with potential corrections from the DM velocity field, we will always consider the DM overdensity field in each combination reducing this number to 17. Moreover, since the velocity dispersion and the vorticity fields are highly correlated, we will also always use the velocity dispersion in the 3D case. Consequently, we limit our study to eight different combinations presented in Table 2. Nevertheless, we have checked that combinations using only DM velocity fields do not lead to satisfactory theoretical predictions.

From each specific association of DM fields, and each Fourier mode k along an LOS, we have estimated the relevant cross-spectra P_{ab} defined in Section 3.1, where a and b refer either to the transmitted flux, the DM overdensity or a specific DM scalar field derived from the DM velocity field. For each mode k, the covariance matrix P_{ab} is Hermitian, and its linear dimension is equal to the total number of fields considered. Examples of cross-power spectra are shown in Fig. 5. We also derived the relevant 1D power spectrum P_k required to computed the covariance Δf_k defined in equation (4). An example of P_k is shown in Fig. 16.

5 CREATING PSEUDO-SPECTRA WITH LYMAS AND LYMAS2

In this section, we apply the LyMAS2 scheme to the DM fields extracted from the HORIZON-NOAGN simulation to generate grids of pseudo-spectra at BOSS resolution. The objective is to recover the 3D Ly α clustering statistics of the 'true' hydro spectra. For a given skewer, we summarize the main steps to follow to produce a corresponding pseudo-spectrum using either LyMAS or LyMAS2:

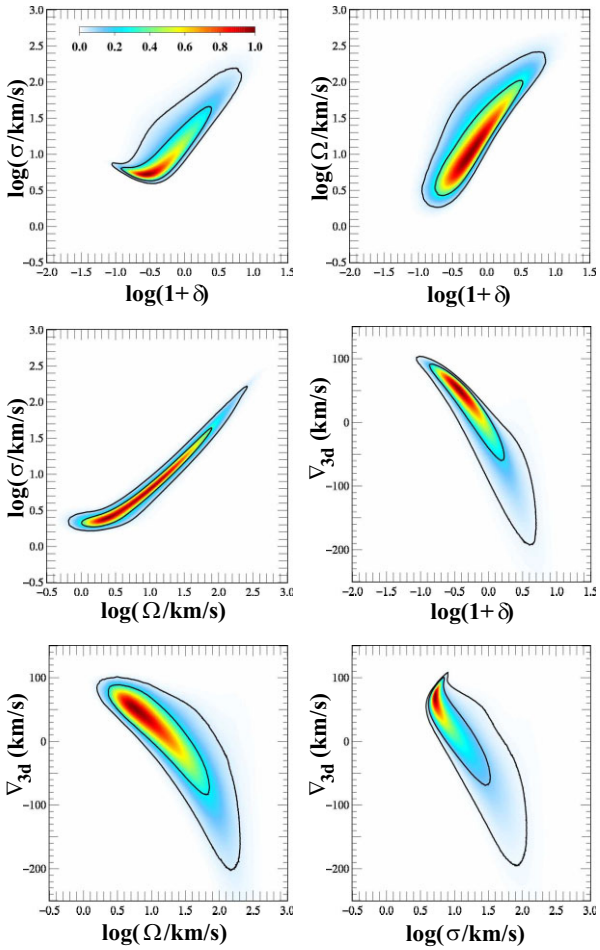

Figure 4. Some correlations between different DM quantities from the Horizon-noAGN simulation at $z = 2.5$. All field have been smoothed to 0.5 Mpc h^{-1} and have been computed in redshift space. Colours show again the density of pixel using normalized values and contour line mark areas enclosing 68.27 and 95.45 per cent.

Table 2. Summary of the different DM field combinations considered in the LyMAS2 scheme.

Name	Field 1	Field 2	Field 3
LyMAS2(ρ)	Overdens.		
LyMAS2(ρ, σ)	Overdens.	Vel. disp.	
LyMAS2(ρ, Ω)	Overdens.	Vorticity	
LyMAS2(ρ, ∇_{1D})	Overdens.	1D div.	
LyMAS2(ρ, ∇_{3D})	Overdens.	3D div.	
LyMAS2(ρ, σ, Ω)	Overdens.	Vel. disp.	Vorticity
LyMAS2(ρ, σ, ∇_{1D})	Overdens.	Vel. disp.	1D div.
LyMAS2(ρ, σ, ∇_{3D})	Overdens.	Vel. disp.	3D div.

The LyMAS scheme:

(1) Extract the smoothed overdensity field ρ for a specific skewer.

(2) Create a realization $G_{per}(x)$ of a 1D Gaussian random field from the 1D power spectrum of the Gaussianized percentile spectra derived from the hydro simulation.

(3) De-Gaussianize $g_{per}(x) = G^{-1}(G_{per})$ to get a realization of a percentile spectrum.

(4) Create a pseudo-spectrum by drawing the flux at each pixel from the location in $P(F|\rho)$, implied by the value of $g_{per}(x)$ (see equation 6 in P14).

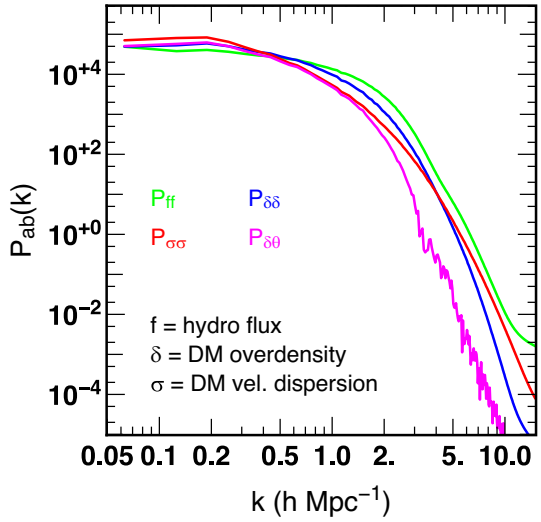

Figure 5. Examples of cross-spectrum $P_{ab}(k) = \langle a_k^* b_k \rangle$, where a and b refer either to the hydro flux or a DM field, derived from the HORIZON-NOAGN simulation at $z = 2.5$. The flux field is 1D smoothed at the BOSS resolution while the DM field are smoothed at 0.5 Mpc h^{-1}.

(5) One full iteration. We first measure the 1D flux power spectrum $P_{ps}(k)$ of the pseudo-spectra created in this way. Then we Fourier transform each pseudo-spectrum and multiply each of its Fourier component by the ratio $[P_F(k)/P_{ps}(k)]^{1/2}$, inverse transform to get the same 1D flux power spectrum than of the true hydro spectra $P_F(k)$ (Weinberg & Cole 1992). The second step of the full iteration is to compute the PDF of the pseudo-spectra after the 1D-P_k re-scaling and then monotonically map the flux value to match the PDF of the true hydro spectra. This full iteration can be repeated several times. However, as we will see, one or two full iterations are enough to get excellent agreement with the 1D power spectrum up to quite high k.

The LyMAS2 scheme:

(1) Extract and Gaussianize the smoothed overdensity field ρ and eventually one or two additional DM velocity fields (e.g. σ) for a specific skewer.

(2) Compute the FFT of each Gaussianized field. This gives new (complex) fields, ρ_k, σ_k, etc.

(3) Compute (in Fourier space) the most probable flux $\bar{f}_k = T_1 \cdot \rho_k + T_2 \cdot \sigma_k + ...$, by applying the relevant filters T_1, T_2, ... (see e.g. equation 12 for the two-fields case).

(4) Generate a 1D Gaussian field of mean 0 and variance defined in equation (4) to get the covariance Δf_k.

(5) After computing the inverse Fourier transform of $\bar{f}_k + \Delta f_k$ to get f, de-Gaussianize to get the pseudo-spectrum: $F = G^{-1}(f)$.

(6) One full iteration. Same procedure as (4) in the LyMAS scheme.

In Fig. 6, we compare the same slice through the hydro flux and through different realizations of LyMAS2 using different combinations of DM fields. The clustering of each pseudo-spectrum is in fair agreement with the clustering of the hydro spectra. Comparing pseudo-spectra and hydro flux along a specific skewer, also shown in Fig. 6, confirms that LyMAS2 correctly models low and high absorptions at good locations, though amplitudes may differ. The second line of Fig. 6 compares three pseudo-spectra generated with LyMAS2 using three different field combinations but using the same seed for the random process to get the variance Δf_k. It's interesting to see that these different pseudo-spectra look also the same, which explains why the slices presented in Fig. 6 are very similar. On the

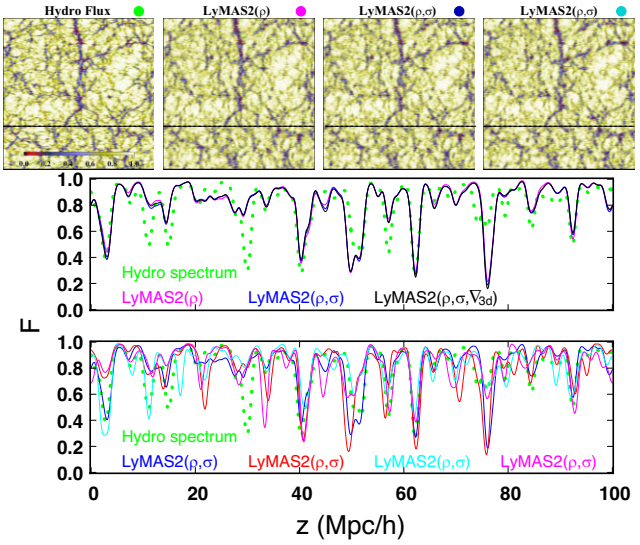

Figure 6. First line: comparison between corresponding slices through the hydro flux (top left-hand panel), 1D smoothed at the BOSS resolution, and through pseudo-spectra generated with LyMAS2 and different associations of DM fields (3D smoothed at 0.5 Mpc h^{-1}). The direction of redshift distortion is horizontal. Flux and DM fields have been extracted from the HORIZON-NOAGN simulation at $z = 2.5$. Visual inspection suggests a very good agreement between the clustering of the hydro flux and those of pseudo-spectra. The two bottom panels show the evolution through a specific LOS (located by the dashed line in the different slices) of the hydro flux (green dotted line) and pseudo-spectra generated with LyMAS2. In the top panel, we use the same seed to randomly generate the covariance from equation (4) for each LyMAS2 realization. We then note that the difference between the different pseudo-spectra is quite weak. In the bottom panel, we consider LyMAS2(ρ, σ) to produce different realizations of pseudo-spectra using this time different seeds to get the covariance. In this case, the difference between pseudo-spectra can be much more pronounced.

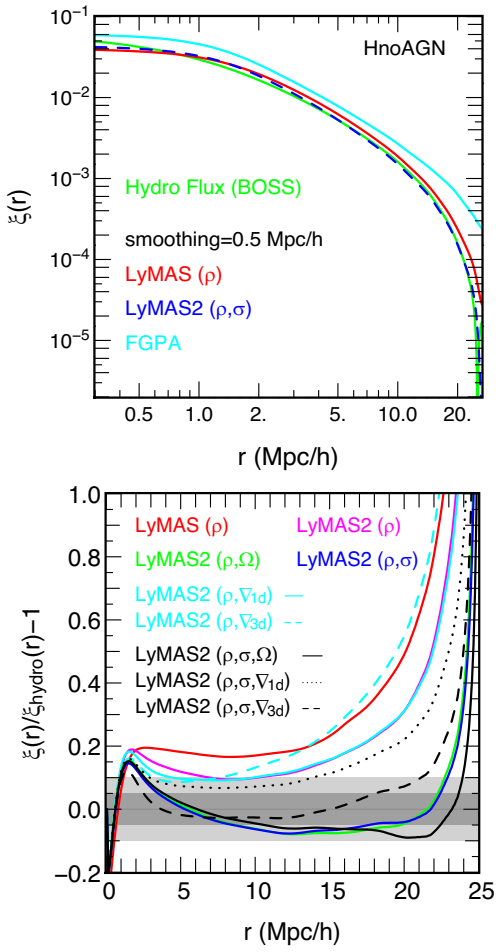

Figure 7. Top panel: the correlation function ξ as a function of the separation r. We present results from the true hydro spectra (green line) and results from the LyMAS (red line) and the new version LyMAS2 considering the DM overdensity and the velocity dispersion fields. The predictions from FGPA (cyan line) are also shown. DM fields are all smoothed at the scale 0.5 Mpc h^{-1}. Bottom panel: the relative difference as respect to the true hydro spectra (i.e. $\xi(r)/\xi_{\rm hydro}(r) - 1$) are shown for different combinations of choices in the Wiener filtering. Here all DM fields have been smoothed to 0.5 Mpc h^{-1}. Compared to the traditional LyMAS scheme (red lines), the new version significantly improves the predictions. In particular, the DM overdensity field associated with the DM velocity dispersion (blue line) or the vorticity (green line) leads to relative difference lower than 10 per cent and close to 5 per cent in most of the range we are interested in. In contrast, using the velocity divergence (cyan lines) does not seem to improve much the results.

contrary, the third line of Fig. 6 shows four different realizations of pseudo-spectrum from LyMAS2 using the DM overdensity and velocity dispersion fields and different seeds to get the covariance Δf_k. In this case, the amplitude of absorptions can be quite different.

In the next sections, we study in more detail the clustering statistics of each catalogue of pseudo-spectra produced with LyMAS2. We aim at recovering three observationally relevant statistics of the transmitted flux: the probability density function (PDF), the LOS power spectrum and the 3D clustering (through the two-point correlation function). As we will see, both LyMAS and LyMAS2 reproduce the PDF of the hydro simulations by construction and nearly reproduce the hydro simulations LOS power spectrum by construction (step 6 above). The power of LyMAS is to produce accurate large-scale 3D clustering while also reproducing these LOS statistics.

5.1 3D-clustering

In order to compare the 3D clustering between the hydro and pseudo-spectra, we rely on the two-point correlation function $\xi(r)$ defined by

$$\xi(r) = \frac{\langle F(x)F(x+r)\rangle}{\langle F\rangle^2} - 1\,, \tag{15}$$

as a function of the separation $\boldsymbol{r}$. To study the effect of redshift distortions, we also consider the two-point correlation function averaged over bin of angle μ defined for a pair of pixels (i, j) by $(r_i - r_j)_\parallel/\boldsymbol{r}$, where $\boldsymbol{r} = |r_i - r_j|$ and $(r_i - r_j)_\parallel$ the component along the LOS.

The top panel of Fig. 7 shows the full two-point correlation functions derived from pseudo-spectra using either the first version of LyMAS (red line) or LyMAS2 using the DM overdensity and velocity dispersion field (blue line). Compared to the results of the hydro flux (green line), one can see that LyMAS2 is significantly improving the predictions that are remarkably close to the hydro spectra results. In order to estimate the precision of these reconstructions, we plot in the bottom panel of Fig. 7 the relative difference, i.e. $\xi/\xi_{\rm hydro} - 1$ for different combinations of DM fields. It appears clearly that LyMAS2 leads in general to much more accurate predictions than LyMAS. Indeed, when considering the DM overdensity field

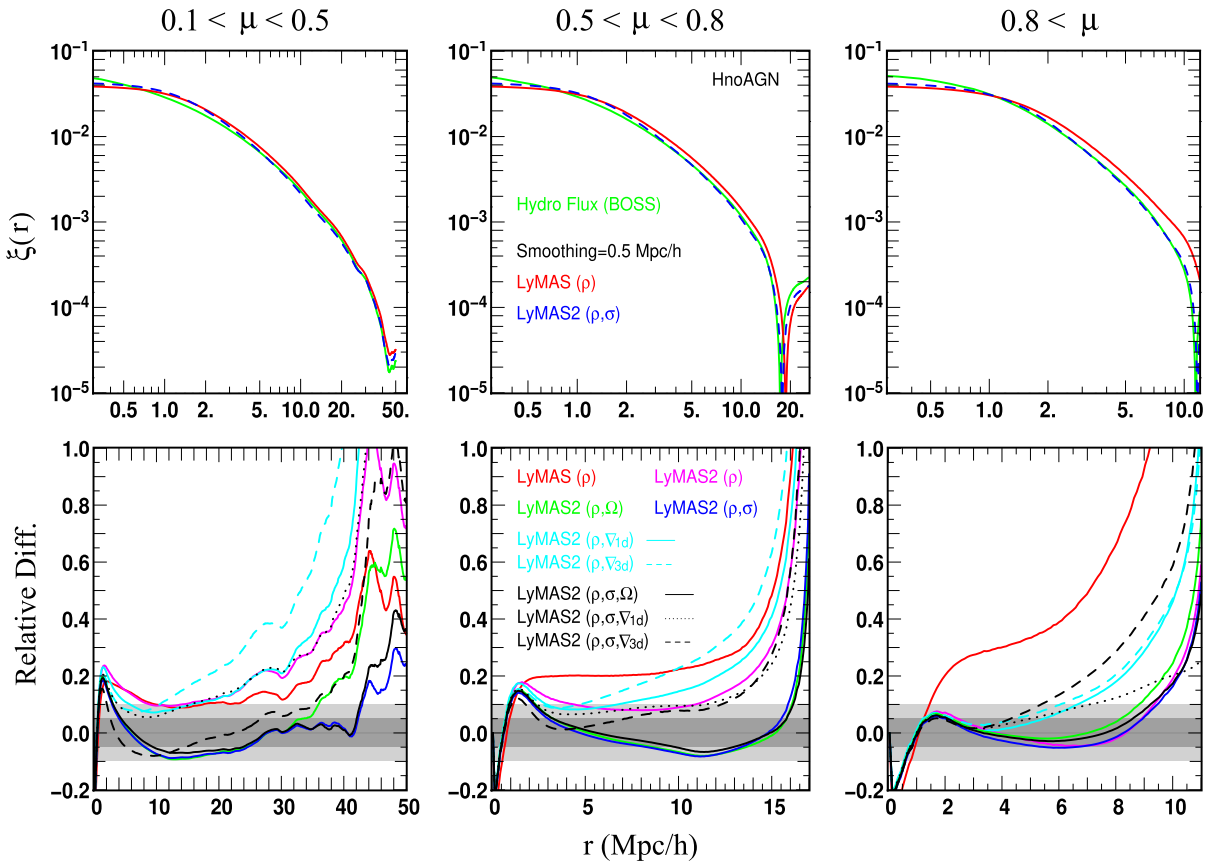

Figure 8. Same as Fig. 7 but with a dependency to range of angles μ, as defined in the text. These plots confirm that the association of the DM overdensity and the velocity dispersion (or vorticity) lead to very accurate predictions (see blue or green lines). The corresponding relative differences shown in the bottom panel are generally within 10 and 5 per cent over a wide range of r. It is also clear that the new LyMAS2 scheme is much more accurate for large angles where the traditional LyMAS leads to errors that grow quickly with the distance r (red lines). DM fields are again smoothed at 0.5 Mpc h^{-1}.

only, LyMAS2(ρ) give slightly better results (magenta line) but the addition of the velocity dispersion lead to errors that are generally lower than 10 per cent and close to 5 per cent (e.g. blue or black lines). Also, similar trends are obtained when the vorticity is taken into account (bottom panel, green line), which is not surprising as these two fields are highly correlated. In contrast, the 1D and 3D velocity divergence, when associated with the DM overdensity only (cyan lines), do not seem to improve much the predictions as respect to the first version of LyMAS. Note that in linear theory, the 3D velocity divergence is fully correlated with the density field and therefore adds no additional information. On the other hand, the vorticity and/or velocity dispersion are sourced by the non-linear evolution of the matter fields and therefore add complementary information on small scales. Finally, we also note that errors are close to 20 per cent at $r \sim$1–2 Mpc h^{-1} probably due to effect of smoothing.

We now investigate how the predictions of the two-point correlation functions vary when considering an angle μ. The trends are presented in Fig. 8 for three ranges of values ($0.1 < \mu < 0.5$, $0.5 < \mu < 0.8$, and $0.8 < \mu$) following P14. The results confirm that LyMAS2 significantly improve the predictions of the Ly α clustering. In particular, some combinations such as (ρ, σ) still lead to errors generally lower than 10 per cent and most of the time close to 5 per cent. We also note that LyMAS2 is particularly efficient for reproducing the correlations along transverse separations or high angles (i.e. $\mu > 0.8$) in which the error is most of the time lower than 5 per cent. The top panels of Fig. 8 indicate again a remarkably good agreement between the two-point correlation functions of the hydro flux and those derived from pseudo-spectra produced from LyMAS2(ρ, σ) even for high angles $\mu > 0.8$) where the previous version of LyMAS is quite inaccurate. For the two largest μ bins, the correlation functions eventually drop rapidly to zero at large r. In this regime, the fractional error in ξ are inevitably large, even though the absolute errors are small. It is evident that LyMAS2 captures the scale of these zero-crossing more accurately than LyMAS.

As a first conclusion, the LyMAS2 scheme is significantly improving the predictions of the Ly α 3D clustering especially when the DM overdensity field is associated with the velocity dispersion or the vorticity field. For the sake of comparison with results presented in the literature, we also compare the 3D power spectrum and corresponding quadropole to monopole ratios in Fig. 9 derived from both the hydro spectra and the pseudo-spectra generated with LyMAS2(ρ, σ). The (monopole) power spectrum is defined in the usual way as $\langle \tilde{F}(\boldsymbol{k})\tilde{F}(\boldsymbol{k}')\rangle = (2\pi)^3 P(k)\delta_{\rm D}(\boldsymbol{k}+\boldsymbol{k}')$, with $\tilde{F}(\boldsymbol{k}) = \int {\rm d}^3x\, F(\boldsymbol{x}){\rm e}^{-i\boldsymbol{k}.\boldsymbol{x}}$. Defined this way, we have the following expression of the variance, $\sigma^2 = \int_0^\infty k^3 P(k){\rm dlog}k/(2\pi^2)$. From Fig. 9, the 3D power spectrum of the hydro spectra can be faithfully recovered from the LyMAS2 simulated spectra up to modes $\sim$2 h Mpc^{-1}. More specifically, the reconstructed Ly α forest power spectrum presents average deviations of $\lesssim$5 per cent up to $k \sim$0.3 h Mpc^{-1}, $\lesssim$10 per cent up to $k \sim$0.4 h Mpc^{-1}, and $\lesssim$20 per cent for modes between 0.4 and 2 h Mpc^{-1}.

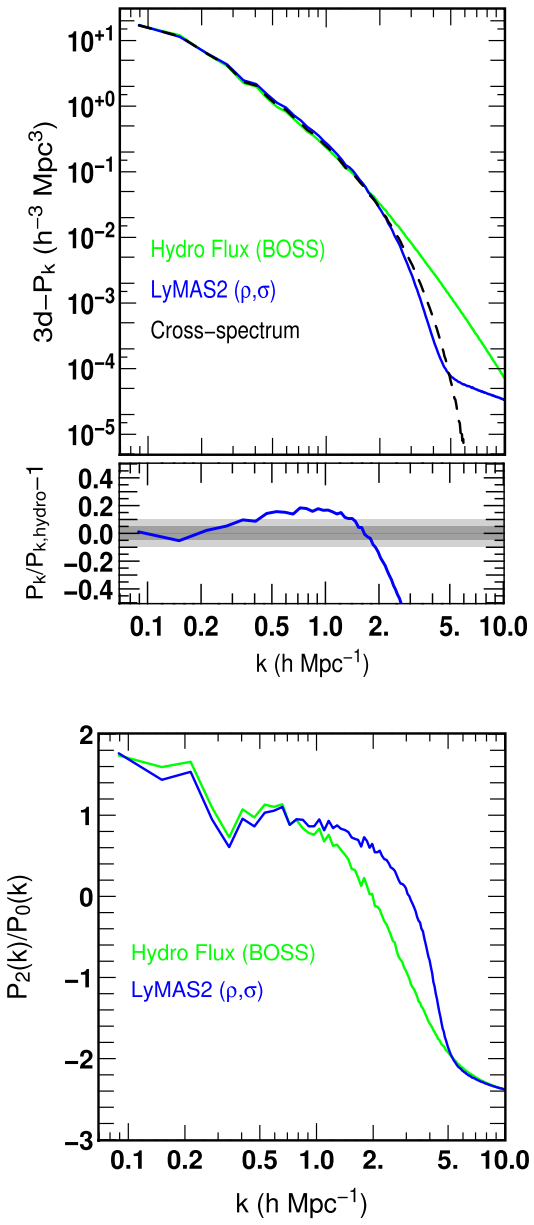

Figure 9. Top panel: the 3D power spectrum derived from the (BOSS) hydro flux (green line) and pseudo-spectra generated with LyMAS2(ρ, σ). All DM fields are smoothed at 0.5 Mpc h^{-1}. The black dashed line represents the cross-spectrum. Bottom panel: the corresponding quadrupole to monopole ratios (same colour code).

For larger modes, however, the predictions are becoming less accurate as separations get lower than the considered smoothing scale (0.5 Mpc h^{-1} here). Indeed, we observe a lack of power at small scales ($2 \leq k \leq 10\,h\,\mathrm{Mpc}^{-1}$) in the 3D power spectrum of LyMAS2 simulated spectra, compared to the hydro power spectrum, which is mainly explained by the fact that the transverse correlations are not accounted for in the Wiener filtering scheme, and in particular transverse fluctuations at small scales are not generated in the present scheme. On the other hand, the absence of correlation, between the stochastic realizations at small scales for each LOS, induces an artificial flattening of the reconstructed power spectrum for modes $k \geq 5\,h\,\mathrm{Mpc}^{-1}$. The ratio of the quadrupole to monopole power is an even stricter test as it traces the anisotropic structure of power in the field, and one can see differences in such a ratio already for modes $k \geq 2\,h\,\mathrm{Mpc}^{-1}$. This test would clearly benefit from accounting for transverse correlations.

It would be interesting to correct this in a forthcoming work though this point is not critical. Indeed the transverse separations of spectra from existing surveys are generally much larger than 1 Mpc h^{-1}, and on these scales the transverse modes are properly reconstructed. Taking into account transverse correlations is straightforward however, and will be worthwhile to generalize this method to emission spectra, for which all transverse scales are important. We will therefore include them in future works.

5.2 1D flux power spectrum along LOS

We also aim at producing catalogues of pseudo-spectra that look like spectra measured by a specific instrument i.e. BOSS in this study. Therefore, the 1D flux power spectrum of each LyMAS mock should be as close as possible to the hydro spectra 1D-P_k. In the following, we only present the results derived from LyMAS2(ρ, σ) as same trends are obtained when considering any other combination of DM fields. Here, the 1D power spectrum is formally defined as $\langle \hat{F}(k)\hat{F}(k')\rangle = (2\pi)P_{1D}(k)\delta_D(k+k')$, where $\hat{F}(k) = \int \mathrm{d}x\, F(x)\mathrm{e}^{-ikx}$ is the 1D Fourier Transform along the LOS.[1] When estimating it, we take FFTs along each LOS and average the result. The expression of the variance is then $\sigma^2 = \int_0^\infty k P_{1d}(k)\mathrm{dlog}(k)/\pi$. Fig. 10 shows the dimensionless 1D power spectrum before power spectrum transformation (red line) and after applying the power spectrum and PDF transformation described in the text (black line). We first note that LyMAS2 without iteration reproduces the 1D power spectrum more accurately than original LyMAS (see fig. 13 of P14). Then as expected, the 1D-P_k transformation leads to same power spectrum as the hydro simulation (blue line), by construction. The second step of the iteration is to re-scale the flux PDF, and this transformation slightly alters the 1D power spectrum. However, as illustrated in the bottom part of Fig. 10, the relative difference is close to 2 per cent up to high values of k (i.e. $k \sim 2\,h\,\mathrm{Mpc}^{-1}$). If one repeats a full iteration a second time, the same accuracy is reached for even higher k values ($\sim 4\,h\,\mathrm{Mpc}^{-1}$).

5.3 One-point PDF of the flux

Since the LyMAS scheme ends after a flux PDF re-scaling (second step of a full iteration), this ensures that the one-point PDF of the hydro flux and the pseudo-spectra match exactly. To illustrate, we present in Fig. 11 the results obtained with LyMAS2(ρ, σ) before (red line) and after (blue line) a full iteration (1D-P_k and PDF re-scaling). Without transformation, the PDF of the pseudo-spectra is already close to the PDF of the hydro spectra (green line). But fractional errors on the PDF can be quite large for low or high values of F. After the 1D-Pk re-scaling, the PDF of the pseudo-spectra has slightly changed and has some non-physical values ($F < 0$ and $F > 1$). But the second step of the iteration corrects the PDF (e.g. transforms these unphysical values into physical ones), to exactly match that of the hydro spectra.

[1]Note that the normalization is reduced by a factor of 4, as compared to the definition used in P14, which relied on a Fourier series in trigonometric functions.

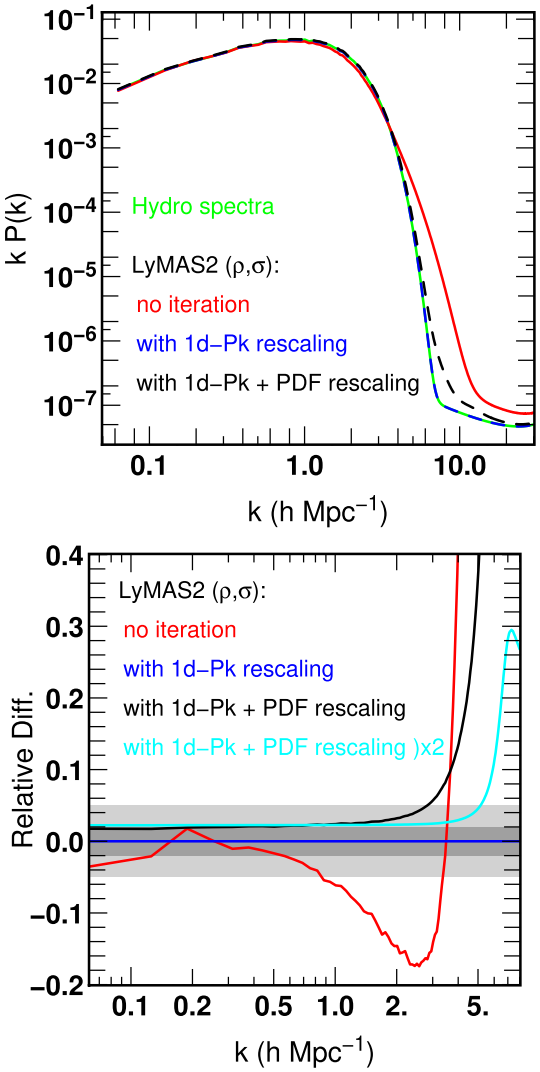

Figure 10. Top panel: the dimensionless 1D power spectrum of the true spectra from the Horizon-noAGN simulation (green line) at $z = 2.5$ and from coherent pseudo-spectra using LyMAS2 considering the DM overdensity and velocity dispersion fields. We show results before (red line) and after (blue and black lines) 1D power spectrum and PDF transformations (described in the text). Bottom panel: the relative difference as respect to the hydro results (i.e. $P_k/P_{k,\,\mathrm{hydro}} - 1$). A full iteration (flux 1D-P_k and PDF re-scaling) permits to recover the hydro power spectrum with an error of ~ 2 per cent over a wide range of k. The light and dark grey shaded areas indicate regions where the error is less than 5 and 2 per cent, respectively.

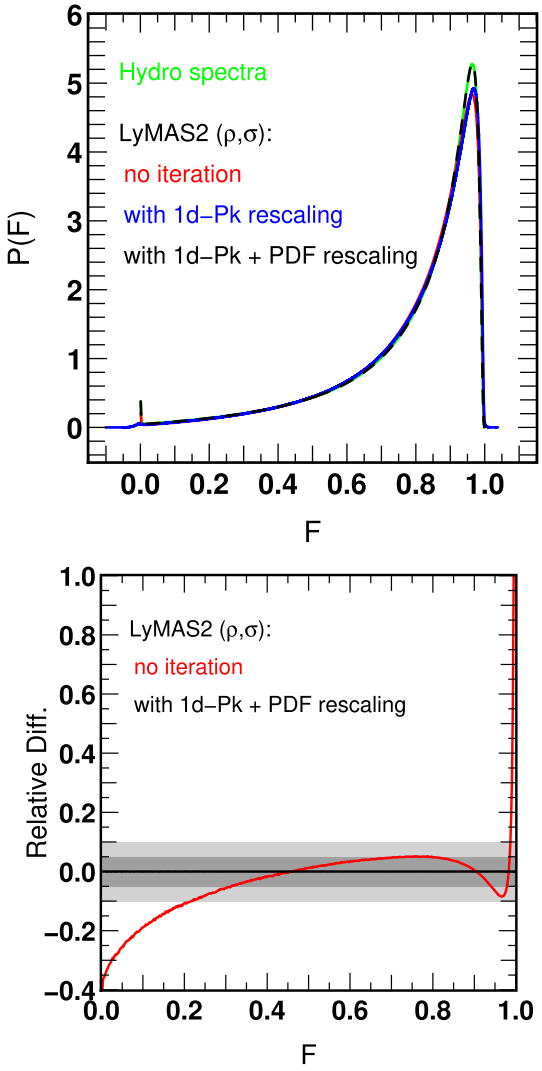

Figure 11. Top panel: the PDF of the true spectra from the HORIZON-NOAGN simulation at $z = 2.5$ (green line) and coherent pseudo-spectra using LyMAS2 considering the DM density and velocity dispersion fields. We show results before (red line) and after (blue and black lines) 1D power spectrum and PDF transformations (described in the text). The 1D-P_k re-scaling can lead to non-physical Flux values (i.e. $F < 0$ or $F > 1$). Bottom panel: the relative difference with respect to the hydro results (i.e. PDF/PDF$_{\mathrm{hydro}} - 1$). The full scheme permits to recover the hydro flux PDF exactly. The light and dark grey shaded areas indicate regions where the error is less than 10 and 5 per cent, respectively.

5.4 Comparison with the FGPA

The FGPA essentially converts DM density into optical depth using a physical model motivated by photoionization equilibrium, assuming that all gas contributing to the Ly α lies on a temperature–density relation $T \propto (\rho_s/\overline{\rho_s})^{\gamma-1}$. The predicted flux is

$$F = Ae^{-(\rho_s/\overline{\rho_s})^{2-0.7(\gamma-1)}}, \quad (16)$$

where $2 - 0.7(\gamma - 1) \approx 0.6$ for the values of γ expected well after reionization (Croft et al. 1998; Weinberg, Katz & Hernquist 1998; McQuinn 2009; Peeples et al. 2010). This relation is reasonable for modelling high-resolution spectra. However, due to existing non-linear relation between flux and optical depth, it does not automatically apply at low resolution (though it omits some physical effects in the high-resolution case). From the HORIZON-NOAGN DM overdensity grid smoothed at 0.5 Mpc h^{-1}, we have first generated 1024×1024 pseudo-spectra using equation (16) by estimating A so that $\langle F \rangle = 0.795$. Then, we 1D smoothed each pseudo-spectrum to BOSS resolution. Similarly to the LyMAS scheme, we end the process by rescaling the flux 1D-P_k and PDF. The correlation function is shown in the top panel of Fig. 7 and is considerably overestimated as respect to the hydro flux with a relative error greater than 50 per cent (omitted in the bottom panel for the sake of clarity). Such a trend is consistent with the results of Sorini et al. (2016), who found that typical relative errors in the 3D power spectrum are ~80 per cent when a DM smoothing scale of 0.4 Mpc h^{-1} is considered. In Appendix B, we will investigate other deterministic mapping than the FGPA. But will we see that the main

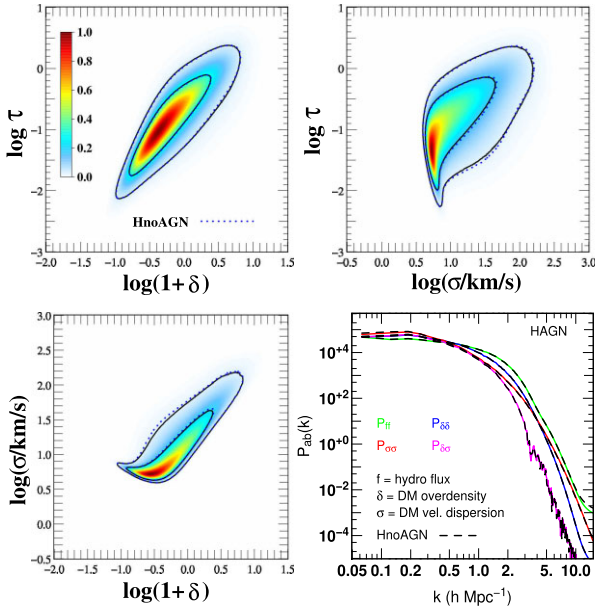

Figure 12. Same as Figs 3 and 4 but using HORIZON-AGN for the calibration. Here we compare the correlations between the optical depth $\tau = -\ln F$ in the hydro spectra (smoothed at the BOSS resolution) and DM quantities smoothed at 0.5 Mpc h^{-1}, namely the overdensity $(1 + \delta)$ and the velocity dispersion (σ) at $z = 2.5$. We also show in the lower right-hand panel, some relevant transfer functions (i.e. cross-spectrum) similarly to Fig. 5. In all panels, the dotted lines correspond to results from the HORIZON-NOAGN simulation. Very similar trends are then obtained when AGN are included or not.

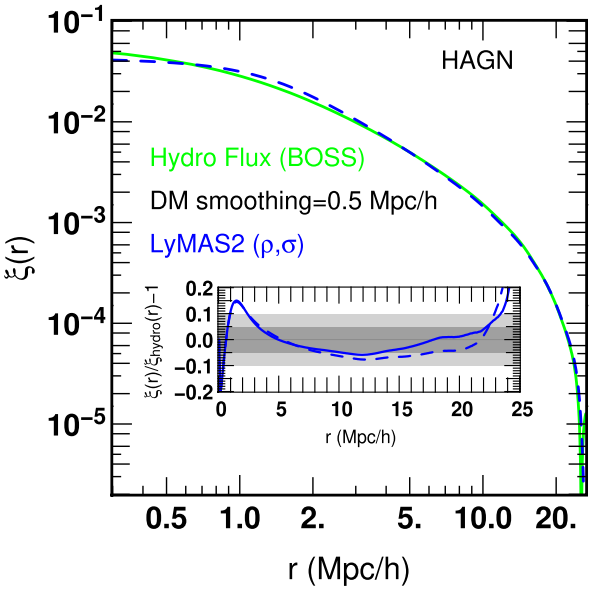

Figure 13. The correlation function ξ as a function of the separation r. We present results from the HORIZON-AGN true hydro spectra (green line) and results from LyMAS2 considering the DM overdensity and the velocity dispersion fields (blue dashed line). DM fields are all smoothed at the scale 0.5 Mpc h^{-1}. The central small panel indicates the relative difference as respect to the true hydro spectra (i.e. $\xi(r)/\xi_{\rm hydro}(r) - 1$). The blue dashed line represents here the result from the HORIZON-NOAGN simulation.

conclusion remains unchanged: deterministic sampling generally tends to significantly overestimate the flux 3D-correlation especially when the DM density is smoothed to scales greater than 0.3 Mpc h^{-1}. Note that a similar trend is obtained when studying the correlation between the Ly α transmitted flux and the mass overdensity (see fig. 1 of Cai et al. 2016).

5.5 HORIZON-AGN versus HORIZON-NOAGN

In this work, we used the HORIZON-NOAGN simulation for the calibration of LyMAS2, mainly to minimize the computational cost as we derived five additional but lower hydrodynamical simulations to study both the robustness of the results (see Appendix A) as well as the effect of cosmic variance (see Section 6.2.1). However, since AGN feedback may induce subtle modifications in the spatial distribution and in the clustering of the Ly α forest, it is important to check if eventual noticeable differences can be seen in the statistics we present so far. For this reason, we have repeated to same and whole analysis but considering this time HORIZON-AGN for the calibration. For instance, we plot in Fig. 12 some relevant scatter plots showing the correlations between the optical depth, the DM overdensity and the DM velocity dispersion, similarly to Figs 3 and 4. We also show some transfer functions (i.e. cross-spectrum) that we compare to results from Fig. 5. In all cases, the statistics have been derived using a DM smoothing of 0.5 Mpc h^{-1}. Compared to results from HORIZON-NOAGN, we found very similar trends when AGN are included. The comparison of the two-point correlation function of the hydro flux and LyMAS2(ρ, σ) pseudo-spectra in Fig. 13 confirms this by suggesting predictions with a very similar accuracy when AGN is included or not. In conclusion, the inclusion of galactic winds does not seem to affect significantly the clustering statistics of the Ly α Forest, given our smoothing scales and targeted accuracy, consistent with results of Bertone & White (2006). Recall that we tuned the UV background in the process of producing the HORIZON-NOAGN hydro flux grid, to get the same mean of the Flux. Thus, this conclusion is not surprising and is in agreement with previous finding (Lochhaas et al. 2016). Above all, this means that the predictions of the 3D clustering from the LyMAS2 scheme keep the same accuracy, AGN feedback included or not in the calibration.

5.6 Influence of redshift distortions?

Our results indicate that the inclusion of a DM velocity field in LyMAS2, especially the velocity dispersion of the vorticity, clearly improves the predictions of the 3D clustering of the pseudo-spectra. This fact might be understood by the existing correlations between the DM overdensity and the velocity fields (see Fig. 4) and adding a velocity term in the scheme may bring additional information. However, since we also model redshift distortions, this may also introduce or enhance existing correlations between the different considered fields. To estimate the importance of the inclusion of redshift distortion in the process, using HORIZON-NOAGN, we have repeated the same analysis in real-space i.e. both hydro flux and DM fields have been generated without redshift distortions. Again, we plot in Fig. 14 some relevant scatter plots showing the correlation between the optical depth, the DM overdensity and the DM velocity dispersion. We also show some transfer functions (i.e. cross-spectrum). In all cases, the statistics have been still derived using a DM smoothing of 0.5 Mpc h^{-1}. Compared to results from HORIZON-NOAGN including redshift distortion, the new scatter plots and transfer functions show significant differences, especially when a DM velocity field is considered. The comparison of the two-points correlation function of the hydro flux and LyMAS2(ρ, σ) pseudo-spectra in Fig. 15 suggests however that errors are still quite

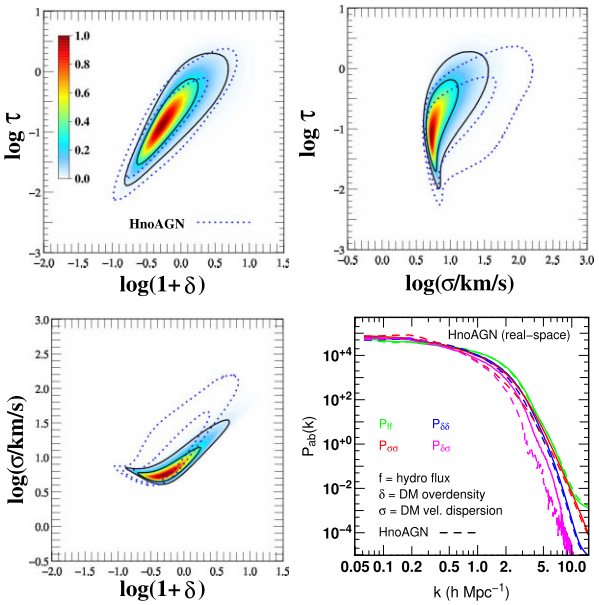

Figure 14. Same as Figs 3 and 4 but using HORIZON-NOAGN without modelling redshift distortions in the LyMAS2 scheme. Here we compare the correlations between the optical depth $\tau = -\ln F$ in the hydro spectra (smoothed at the BOSS resolution) and DM quantities smoothed at 0.5 Mpc h^{-1}, namely the overdensity $(1 + \delta)$ and the velocity dispersion (σ) at $z = 2.5$. We also show in the lower right-hand panel, some relevant transfer functions (i.e. cross-spectrum) similarly to Fig. 5. In all panels, the dotted lines correspond to results from the HORIZON-NOAGN simulation with redshift distortions.

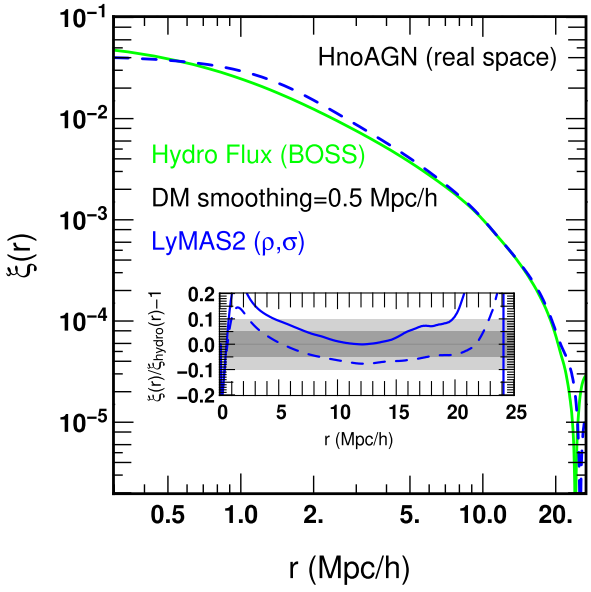

Figure 15. The correlation function ξ as a function of the separation r. We present results from the HORIZON-NOAGN (NO redshift distortions) true hydro spectra (green line) and results from LyMAS2 considering the DM overdensity and the velocity dispersion fields (blue dashed line). DM fields are all smoothed at the scale 0.5 Mpc h^{-1} and are generated in real space. The central small panel indicates the relative difference with respect to the true hydro spectra (i.e. $\xi(r)/\xi_{\rm hydro}(r) - 1$). The blue dashed line represents here the result from the HORIZON-NOAGN simulation including redshift distortions.

low, but a bit higher compared to the relative errors obtained when including redshift distortion. Then, it appears that the inclusion of redshift distortion seems to slightly improve the predictions of the Ly α clustering statistics.

6 APPLICATION TO LARGE COSMOLOGICAL DM SIMULATIONS

6.1 Simulations of 1.0 and 1.5 Gpc h^{-1} boxside

In this section we apply our LyMAS2 scheme to large cosmological DM simulations to produce ensembles of BOSS pseudo-spectra. We first ran five cosmological N-body simulations using GADGET2 (Springel 2005), with a box length of 1.0 Gpc h^{-1} with random initial conditions and using the same cosmological parameters as HORIZON-NOAGN. We additionally run one simulation with a higher volume, namely $(1.5\ {\rm Gpc}\ h)^{-3}$. As we discuss in detail in Section 6.2.2, these latter two values have been chosen to estimate the performances of LyMAS when using DM smoothing scales of 0.5 (fiducial) and 1.0 Mpc h^{-1}, respectively. In each simulation, the adopted value of the Plummer-equivalent force softening is 5 per cent of the mean inter-particle distance (24.4 and 36.6 kpc h^{-1} for the 1.0 and 1.5 Gpc h^{-1} boxside, respectively) and kept constant in comoving units.

From each cosmological simulation, the corresponding DM density and velocity dispersion fields are computed and sampled on grids of 4096^3 pixels. This allows us to smooth each 1.0 Gpc h^{-1} field to 0.5 Mpc h^{-1} and each 1.5 Gpc h^{-1} one to 1.0 Mpc h^{-1}. According to Section 4 and Appendix A, the combination of the DM overdensity and velocity dispersion fields leads to accurate and robust Ly α clustering predictions. We therefore produce our fiducial large BOSS pseudo-spectra with LyMAS2(ρ, σ). Note that we choose the velocity dispersion field instead of the vorticity mainly for practical reasons, as the computational and memory costs to compute the latter on a large regular grid is much higher.

Once the different DM fields are extracted and smoothed to the appropriate scales, the last inputs we need are the relevant transfer functions T defined in equation (10) whose detailed expressions can be found in equations (11) and (12) for the 1D or 2D case, respectively. We also need the corresponding 1D power spectrum P_k to generate the covariance Δf_k at the considered boxside. Since the calibrations are derived from the HORIZON-NOAGN simulation, one potential issue arises from the hydrodynamical box being much smaller than the large DM simulations, so that lower modes are not represented. For the missing modes ($k \leq 2\pi/100$), we have extrapolated the values of T_1, T_2, and P_k, while in the common k range, we have proceeded with interpolations. As an illustration, Fig. 16 shows the 1D power spectrum required to compute the covariance Δf_k when considering the DM density and velocity fields extracted from the 100 Mpc h^{-1} hydrodynamical simulation as well as the resulting 1D power spectra when considering a 1.0 or 1.5 Gpc h^{-1} boxside.

Fig. 17 illustrates a reconstruction of pseudo-spectra from a given slice of 4096×4096 pixels through a 1 Gpc h^{-1} box simulation. It appears clearly that the 2D clustering of the pseudo-spectra agrees very well with the clustering of the DM overdensity field. Another visual inspection of an individual skewer also shows that peaks of density match with high absorption. It is also interesting to see that the specific skewer shown in Fig. 17 has in its centre a large absorption that corresponds to a large- and high-density region. Note that the study of groups of so-called 'Coherently Strong Ly α Absorption' (CoSLA) systems imprinted in the absorption spectra of a number of quasars (from e.g. BOSS) is of particular interest, as they can potentially detect and trace high redshift proto-clusters (see e.g.

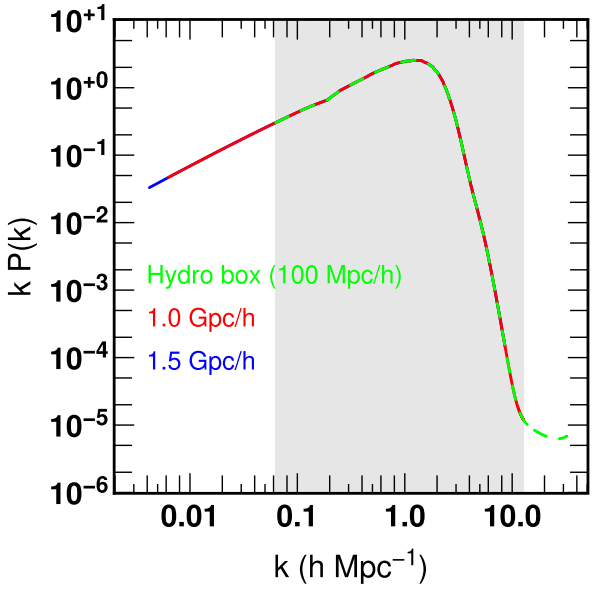

Figure 16. An example of dimensionless 1D-P_k required to compute the covariance Δf_k (see equation 4), derived from the HORIZON-NOAGN simulation at $z = 2.5$ and considering the overdensity and velocity dispersion in the Wiener filtering (green dashed line). The resolution of the flux and DM grids are 1024^3. The red and blue lines are interpolations and extrapolation of the green line to construct the corresponding 1D-P_k for larger 1.0 and 1.5 Gpc h^{-1} boxes (using grids of resolution 4096^3). The grey shaded area indicates the common k range between the 100 Mpc h^{-1} and 1 Gpc h^{-1} boxes.

Francis & Hewett 1993; Cai et al. 2016; Lee et al. 2018; Shi et al. 2021).

Fig. 18 shows the dimensionless 1D power spectrum of the pseudo-spectra from a 1 Gpc h^{-1} simulation before and after iterations. First, in the common k-range area between the hydro and the pseudo-spectra, we find similar trends to those obtained when applying LyMAS2 to the HORIZON-NOAGN simulation (see Fig. 10). For instance, after two full iterations, the relative difference is close to 2 per cent even for high values of k ($\sim$4 h Mpc^{-1}), and similar results are obtained for the 1.5 Gpc h^{-1} pseudo-spectra. For lower values of k ($k \leq 2\pi/100 \sim 0.0628\,h$Mpc^{-1}), the power spectrum seems to have a natural and consistent extension from the hydro spectra power spectrum. Note also that the highest values of k for the 1.0 and 1.5 Gpc h^{-1} boxside simulations and grid of resolution 4096^3 are, respectively, 12.87 and 8.58 Mpc^{-1}, lower than $(2\pi/100) \times 512 \sim 32.17\,h$Mpc^{-1} for the hydro box. But the power at these high values is negligible, and missing them in the calculations will not have a noticeable impact on spectra. Also, since a full iteration ends with a flux PDF re-scaling, this ensures exactly match to the PDF of the hydro flux.

Fig. 19 shows the two-point correlation functions derived from several large-scale pseudo-spectra. In particular, we show the predictions from the first version of LyMAS (red lines) and those obtained from LyMAS2 using the DM density field only (black lines) and with additional velocity dispersion field (blue lines). We also add the predictions derived from the 1.5 Gpc h^{-1} simulation (magenta lines), using again LyMAS2(ρ, σ). These plots confirm first that the traditional LyMAS (red lines) tends to overestimate the correlations and this trend is more pronounced when considering high angles ($\mu > 0.8$), as already noted in Section 4. The result is quite similar with the LyMAS2 scheme when considering the DM overdensity only. However, the difference from LyMAS is more and more noticeable as μ increases. These results are again consistent with those presented in Section 4. The difference becomes even stronger when adding the DM velocity dispersion field. In this case, LyMAS2(ρ, σ) tends to significantly reduce the correlations and most probably lead to more reliable predictions. In the range $2 \leq r \leq 10$ Mpc h^{-1}, the correlations are very close to those of the hydro simulation. It is also impressive that the 1.5 Gpc h^{-1} mock generated with LyMAS2(ρ, σ) leads to very similar trends (for separations $r \geq 2$ Mpc h^{-1}), though the DM fields are now smoothed to 1.0 Mpc h^{-1}. This success is consistent with the results presented in the Appendix A, where we compare the performance of LyMAS2 using different DM smoothing scales. This robustness is one of the key improvements accomplished with LyMAS2.

Finally, we show in Fig. 20 the two-point correlation function averaged from five different realizations of 1 Gpc h^{-1} Ly α pseudo-spectra obtained by applying LyMAS2(ρ, σ) to different DM cosmological simulations, at $z = 2.5$. The plots show clear features of BAO at $r \sim 105$ Mpc h^{-1} and variations with respect to the angle μ, consistent with observational trends (see e.g. du Mas des Bourboux et al. 2020) This illustrates the ability of LyMAS to properly describe redshift distortions and to model realistic large BOSS Ly α forest spectra catalogues.

6.2 Potential limitations of the method

6.2.1 Effect of cosmic variance?

One potential limitation in the LyMAS scheme is to use an unique hydro simulation to generate the calibration. In other words, we assume this hydro simulation to be fairly representative of the underlying statistics of many simulations that have $\geq$1000 times larger volumes. This makes the resulting large mocks potentially affected by the cosmic variance. In order to estimate this, we have considered our five lower resolution hydro simulations presented in Appendix A, originally produced to estimated the robustness of the LyMAS2 predictions. Here we make good use to estimate the effect of cosmic variance by applying each of the five calibration sets to the $(1.5\ \text{Gpc}\ h^{-1})^3$ DM overdensity and velocity grids (of dimension 4096^3 each) that we used in Section 6.1. In particular, we have computed the two-point correlation function averaged from these five realizations and shown in Fig. 21. We note that the dispersion tends to be higher for separations between 25 and 100 Mpc h^{-1}, which makes sense as this corresponds to the scales probed by the reference hydro simulation of boxside 100 Mpc h^{-1}. We also note that this dispersion tends to be higher for increasing values of the angle μ.

6.2.2 Computational limitations

As most of the methods presented in the literature to produce large Ly α mock catalogues, the LyMAS2 scheme can be divided into two main operations. One the one hand, one needs to generate at a considered smoothing scales, a DM overdensity field and eventually associated velocity fields. This task is generally done using N-body simulations or lognormal density fields created from Gaussian initial conditions (e.g. Gnedin & Hui 1996; Bi & Davidsen 1997). On the other hand, one has to 'paint' the Ly α absorptions from any los using relevant calibrations or recipes. This latter part is pretty fast in LyMAS since one can treat any los individually and therefore, the algorithm can be easily and optimally parallelized (with openMP for instance). To give an order of magnitude, to

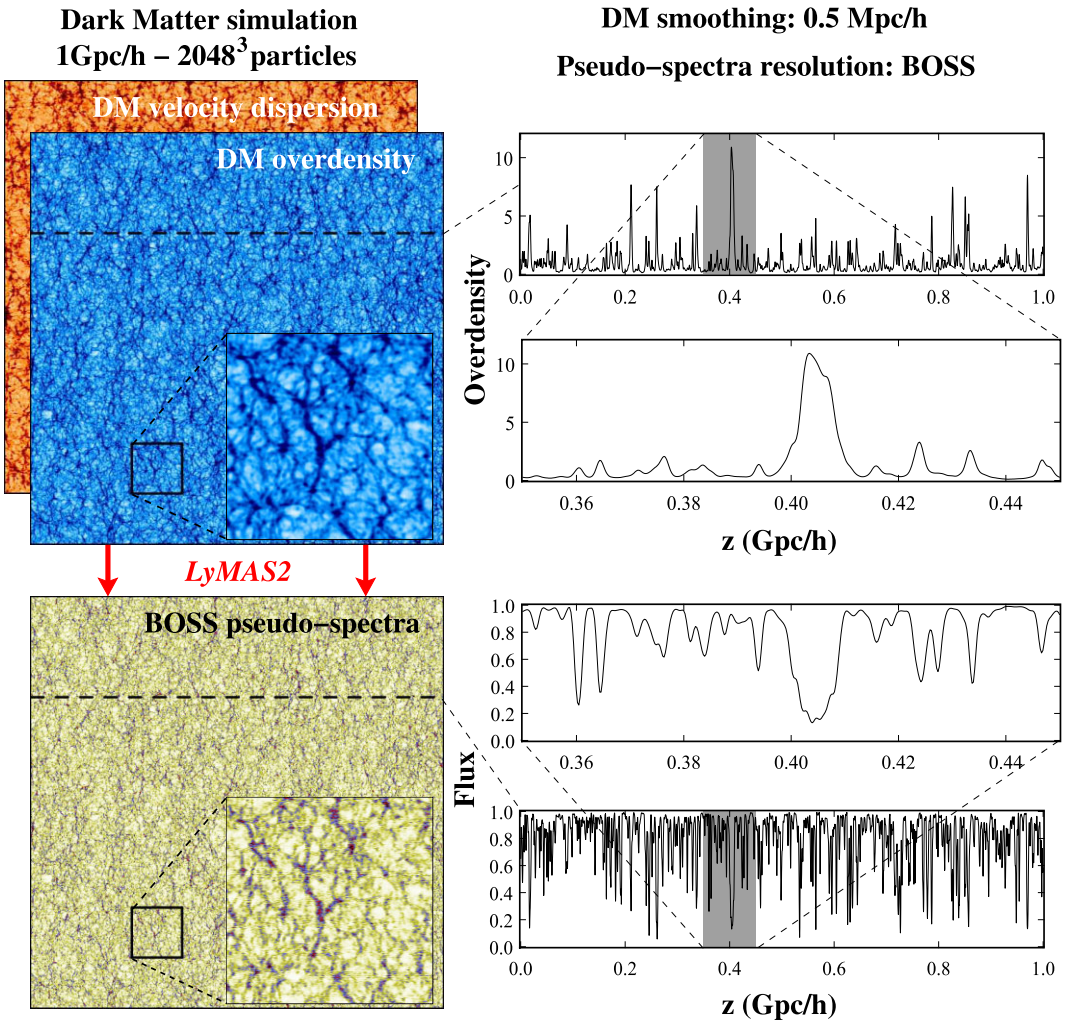

Figure 17. Application of the full LyMAS2 scheme to a large DM simulation (1 Gpc h^{-1} – 2048^3 particles) at $z = 2.5$. From the DM overdensity and velocity dispersion fields, both sampled on 4096^3 regular grids and smoothed at 0.5 Mpc h^{-1}, we derive pseudo-sepctra at the BOSS resolution using the full LyMAS2 scheme. The left-hand part of the figure shows corresponding slices that suggest a fair agreement between the clustering of the DM density and the pseudo-spectra. The right-hand part shows an individual skewer confirming that DM density peaks are associated with high flux absorption.

create a $\geq$1 Gpc h^{-1} boxside Ly α mock presented in Section 6.1, using 32 CPU, only $\sim$11 h are required to generate 4096×4096 BOSS spectra of resolution 4096 and subsequent 1D-P_k and flux PDF rescaling (namely to achieve the six steps of the LyMAS2 scheme presented in Section 5). The main limitation of LyMAS2 is however the ability to generate the DM overdensity or a velocity field at the appropriated smoothing scales (i.e. 0.5 and 1.0 Mpc h^{-1} in our study). Indeed, the larger the boxside of the simulation, the bigger the required dimension of the grid to sample the DM fields. For instance, a simulation box of side 1.0 or 1.5 Gpc h^{-1} can be smoothed at the scale of 0.5 and 1.0 Mpc h^{-1}, respectively, if a grid of 4096^3 pixels is considered. In these cases, the size of individual pixel is, respectively, 0.244 and 0.366 Mpc h^{-1}, which is acceptable, though slightly borderline, to produce the smoothing operation. Moreover, to generate one specific DM field, we used a sophisticated scheme, SmoothDens5, presented in the Appendix C. Although SmoothDens5 has been optimized, it needs at least 1-TB RAM and 50 h (using 64 CPU) to treat and produce a single DM field, sampled on a regular grid of 4096^3, and from a DM simulation using 2048^3 particles.

Technically, it is then rather feasible to generate massive set of mocks if we limit the studied boxside to 1.5 Gpc h^{-1}. Beyond this value, the computational costs and memory requirement is becoming an issue. It would be definitely worth exploring in near future alternative methods to reduce such costs (e.g. Cell-in-Cloud,...) while not altering the accuracy of the predictions. Moreover, although this is a general issue for all the methods based on DM fields described by N-body simulations, N-body simulations can also become too computationally expensive and time-consuming. Here also alternative methods do exist to obtain the DM fields using cheap approximate methods (e.g. LPT, 2LPT, etc.). However, these are typically not able to produce a very accurate velocity field and this may alter the accuracy of the present LyMAS scheme. Such investigations are beyond the scope of this paper and will be considered in the next analysis.

7 CONCLUSIONS

We have introduced LyMAS2, an improved version of the LyMAS scheme (P14). In this new version, we have used the HORIZON-NOAGN (Peirani et al. 2017) simulation to characterize the relevant cross-correlations between the transmitted flux and the different DM fields. In particular, we have considered not only the DM overdensity

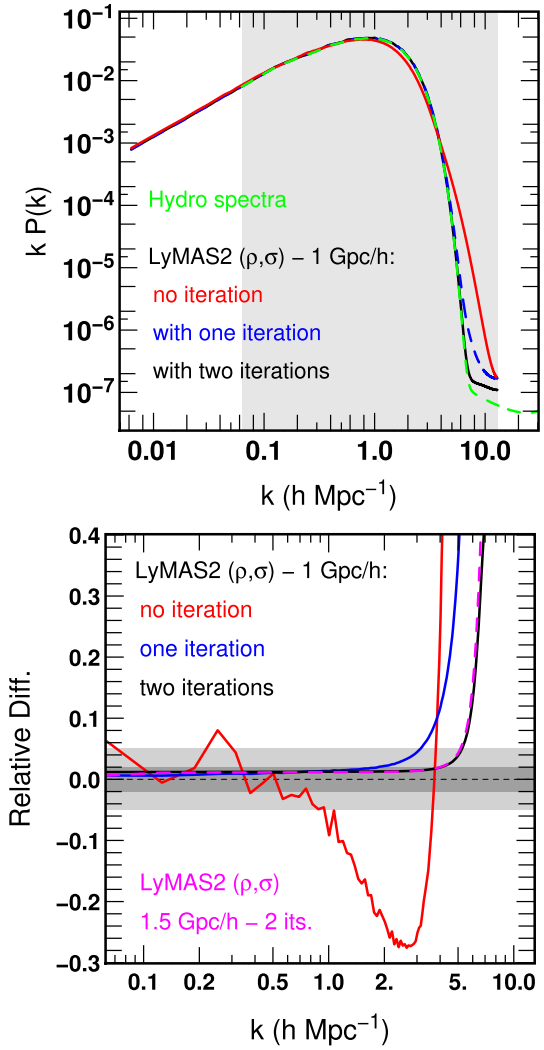

Figure 18. Top panel: the dimensionless, redshift-space, 1D power spectrum of pseudo and true spectra from the HORIZON-NOAGN simulation (green line) and from coherent pseudo-spectra using LyMAS2 considering the DM density and the velocity dispersion fields from a 1 Gpc h^{-1} boxside N-body simulation. Here again, we show results before (red line) and after (blue line) one full iteration (i.e. flux 1D power spectrum and PDF transformations). We also show the results when repeating a second iteration (black line). The grey shape defines the common k-range between hydro spectra and pseudo-spectra. Bottom panel: the relative difference with respect to the hydro results (i.e. $P_k / P_{k,\mathrm{hydro}} - 1$). The light grey and dark grey bands define regions where the error is less than 5 and 2 per cent, respectively.

but also specific DM velocity fields (i.e. velocity dispersion, vorticity, 1D and 3D divergence) and used Wiener filtering to generate the specific calibrations. LyMAS2 shares the same philosophy as LyMAS that flux correlations are mainly driven by the correlations of the underlying DM (over)density, and it uses additional information from the DM velocity correlations to refine the theoretical predictions. In a second step, we have applied LyMAS2 to DM fields extracted from the hydrodynamical or large DM-only simulations to create large ensembles of pseudo-spectra with redshift distortions, at $z = 2.5$ and at the BOSS resolution. Throughout the analysis, we use a DM smoothing of 0.5 Mpc h^{-1} to derive the main trends and results. Our main conclusions can be summarized as follows:

(i) LyMAS2 greatly improves the predictions for flux statistics of the 3D Ly α forest on small and large scales. More specifically, we found that the DM overdensity combined with the DM velocity dispersion (or the vorticity) recovers the two-point correlation functions of the (reference) hydro flux within 10 per cent and (most of the time within 5 per cent) even when high angles are considered. This is a major improvement with respect to the original version of LyMAS, which is rather inaccurate in predicting the Ly α correlations for large separations and high angles. furthermore, we found that the reconstructed Ly α forest power spectrum presents average deviations of $\lesssim$5 per cent up to $k \sim 0.3\,h\,\mathrm{Mpc}^{-1}$, $\lesssim$10 per cent up to $k \sim 0.4\,h\,\mathrm{Mpc}^{-1}$ and $\lesssim$20 per cent for modes between 0.4 and $2\,h\,\mathrm{Mpc}^{-1}$. For larger modes, however, the predictions are becoming less accurate as separations get close or lower than the considered smoothing scale (typically 0.5 Mpc h^{-1}).

(ii) Like LyMAS, LyMAS2 reproduces the one-point PDF of the flux from the calibrating hydro simulation exactly, by construction. It also reproduces the 1D (LOS) power spectrum with en error of about 2 per cent up high k values. The LyMAS2 pseudo-spectra therefore have realistic observable properties on small scales while also having accurate large-scale 3D clustering when applied to a large-volume DM-only simulation.

(iii) The trends derived from five different and slightly lower resolution hydrodynamical simulations are consistent with those obtained from the fiducial HORIZON-NOAGN simulation. This suggests that the results presented in this study are robust. Moreover, this allows us to estimate error bars on the two-point correlations functions, which are generally low.

(iv) We have considered three different DM smoothing scales (0.3, 0.5, and 1.0 Mpc h^{-1}) and found similar trends in the flux clustering predictions. It is encouraging that a DM smoothing of 1.0 Mpc h^{-1} still leads to very accurate predictions, especially in the two-point correlation functions even at high angles and large separation. Indeed, the errors are typically lower than 5 per cent, whereas they are generally higher than 30 per cent with the original version of LyMAS.

(v) LyMAS2 applied to large DM cosmological simulations of boxside either 1.0 or 1.5 Gpc h^{-1} indicates that the predicted flux statistics follow the same trends obtained from the (100 Mpc h^{-1}) HORIZON-NOAGN DM fields. Indeed, we found again that the first version of LyMAS tends to overestimate the flux correlations at large separations and/or at high angles. On the contrary, LyMAS2 using for instance the DM overdensity and the velocity dispersion clearly reduces the two-point correlation functions to lead to more reliable and accurate predictions. Moreover LyMAS2 adequately models large-scale Ly α absorptions systems that correspond to massive overdensity regions. It is also worth mentioning that these set of mocks were already used to asses the ability to recover the connectivity and clustering properties of critical points of the reconstructed large-scale structure from Ly α tomography in the context of a realistic quasar survey configuration such as WEAVE-QSO (Kraljic et al. 2022).

(vi) Deterministic mappings such as the Fluctuating Gunn–Peterson Approximation tend to considerably overestimate the 3D flux correlations especially at large separation or when high angles are considered.

LyMAS2 offers a sophisticated tool to accurately model and predict large-scale Ly α forest 3D statistics. This opens new opportunities to improve diversified studies such as Ly α forest cross-correlation (e.g. Lochhaas et al. 2016), two-point correlations or three-point correlations analysis (e.g. Tie et al. 2019) or BAO feature predictions. Moreover, large Ly α catalogues produced with LyMAS2 can be used to characterize massive overdensity regions such as proto-clusters through groups of coherent large absorptions analysis (Cai et al. 2016; Lee et al. 2018; Shi et al. 2021). Compared

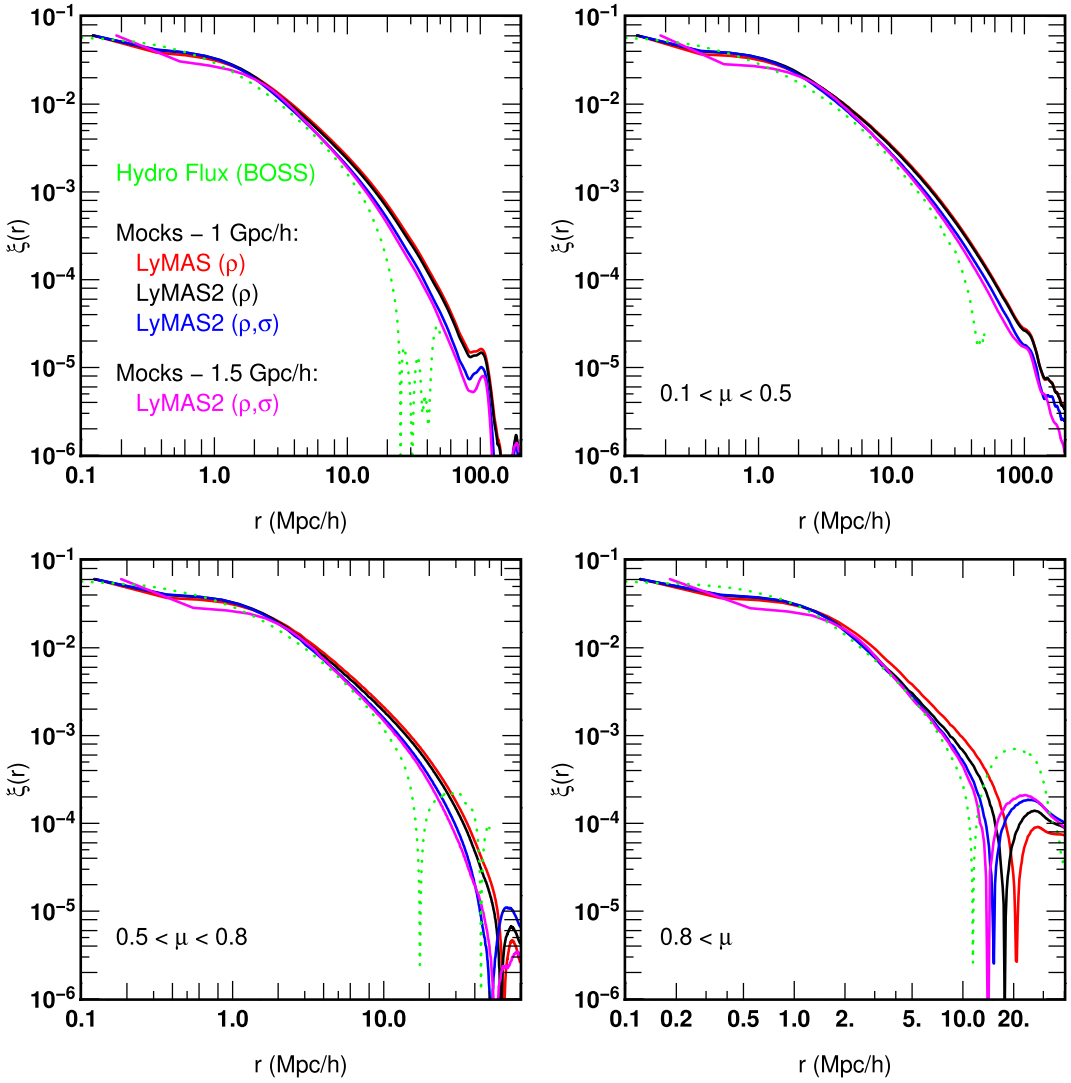

Figure 19. Top left-hand panel: the two-point correlation functions of the flux derived from the 100 Mpc h^{-1} HORIZON-NOAGN simulation (green dotted line) and 1 Gpc h^{-1} pseudo-spectra using the full LyMAS (red line) and LyMAS2 scheme using the DM overdensity field only (black line) or combined with the DM velocity field (blue line). We also show the two-point correlation function derived from 1.5 Gpc h^{-1} pseudo-spectra produced with LyMAS2(ρ, σ) (magenta line). In the other panels, we show the corresponding flux correlation functions averaged over bins of angle μ, as labelled. LyMAS tends to overestimate the correlation especially for large separations and high angles. LyMAS2(ρ, σ) tends to significantly reduce such correlations, which suggests more reliable predictions.

to previous work, we recall that the main objective of LyMAS is to create large Ly α mocks for a specific instrument (here BOSS) with 3D flux statistics as close as possible to those that would be obtained from a very large volume (but computationally intractable) hydrodynamical simulation.

The Iteratively Matched Statistics (IMS) developed by Sorini et al. (2016) does not present such predictions and limits their analysis to small simulation boxes ($\leq$100 Mpc h^{-1}). However, when comparing the flux PDF and 1D power spectrum, LyMAS2 and 1D-IMS (see introduction) lead to similar performances: The 1D-IMS scheme perfectly reproduces these statistics, while errors of $\sim$2 per cent are obtained with LyMAS2 for the flux 1D-P_k. Regarding the 3D-IMS scheme, errors are much higher, of the order of 15 and 20 per cent, respectively, for the flux PDF and 1D power spectrum. As far as the 3D flux statistics are concerned, at a DM smoothing of 0.4 Mpc h^{-1}, the 1D-IMS and 3D-IMS present errors of 20 per cent and 10–20 per cent (for a DM smoothing of 0.4 Mpc h^{-1}), respectively, regarding the reconstruction of the power spectrum. In this study, LyMAS2 mainly considers a DM smoothing of 0.5 Mpc h^{-1}, which leads to errors generally lower than 5 per cent for the two-point correlation functions. Again, it is worth mentioning that similar (low) errors are also obtained with LyMAS2 when considering a DM smoothing of 1.0 Mpc h^{-1}. It would be then interesting to compare the performance of the 1D and 3D-IMS scheme at this specific smoothing scale in the perspective of creating large ($\geq$ 1.0 Gpc h^{-1}) Ly α mocks.

Recently, Harrington et al. (2022) have trained a convolutional neural network from hydrodynamical simulations of side 20 Mpc h^{-1} to predict both the density, the temperature and the velocity fields. This method is quite flexible and the predictions of the flux PDF and 1D power spectrum (i.e. within $\sim$5 per cent up to $k \sim 10$ Mpc h^{-1}) are promising and more accurate than the FGPA. Note that in a companion paper (Horowitz et al. 2021), convolutional neural networks have also been used to synthesize hydrodynamic

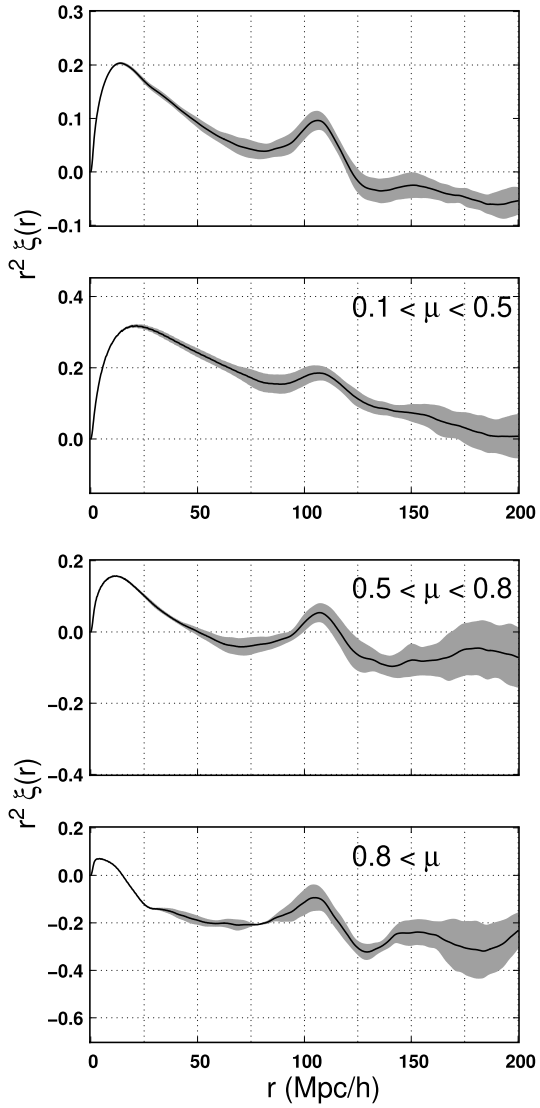

Figure 20. The two-point correlation function averaged from five different realizations of 1 Gpc h^{-1} Ly α pseudo-spectra obtained by applying LyMAS2(ρ, σ) to DM cosmological simulations. Here we use the calibrations obtained from HORIZON-NOAGN. The shaded areas represent the error on the mean (rms).

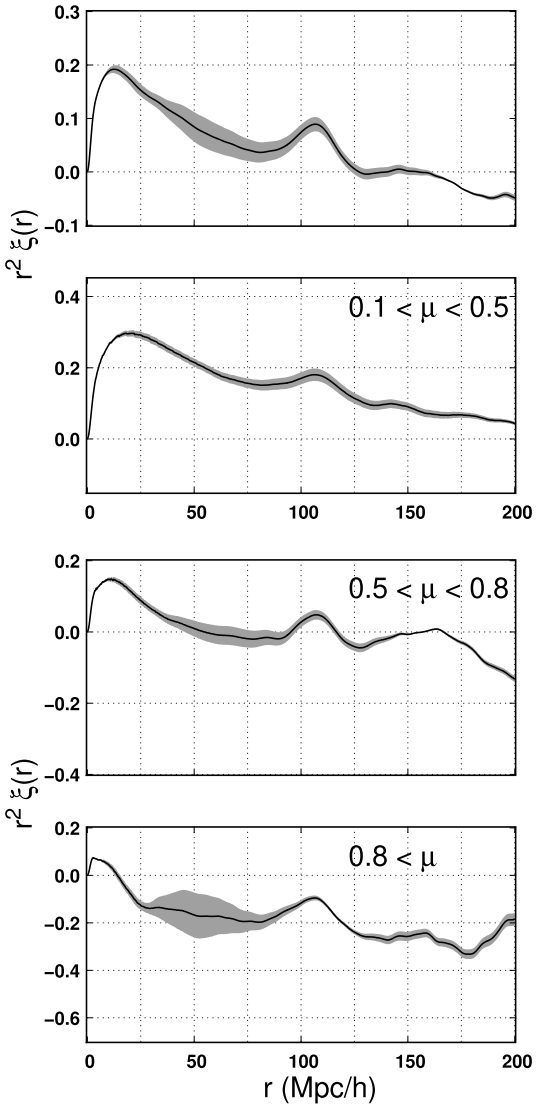

Figure 21. The two-point correlation function averaged from five different realizations of 1.5 Gpc h^{-1} Ly α pseudo-spectra obtained by applying LyMAS2(ρ, σ) to DM cosmological simulations and using five different calibrations. To this regard, we use five lower resolution hydro simulations presented in Appendix A. The shaded areas represent the error on the mean (rms).

fields conditioned on DM fields from *N*-body simulations, which might be very useful for the rapid generation of mocks. Similarly, Sinigaglia et al. (2022) has developed a new physically motivated supervised machine learning method (HYDRO-BAM) from a reference hydrodynamical simulation of comoving side 100 Mpc h^{-1}. The PDF, 3D power spectrum and bi-spectra can be reconstructed with error of a few per cent up to modes $k = 0.9$ Mpc h^{-1}. It would be interesting to see how this promising approach performs when considering smoothed spectra and larger boxes.

Improvements can still be done in the LyMAS scheme. For instance, one main assumption is to consider that the transverse correlations are mainly driven by the effect of DM smoothing. In this study, we stress again that all the approach is based on creating pseudo-spectra individually and independently from each other. Because the draws of Δf_k are independent on each LOS, spectra at small but non-zero transverse separations can look quite different on small scales. Since the predictions on the clustering of the flux are already very accurate with LyMAS2, we have not considered the same approach in volume. This would take into account transverse correlations between LOS that have been neglected in this work: instead of predicting the flux from DM fields independently for each LOS, one would predict the entire cube of flux from the DM field cubes, using the full 3D covariance structure. One would still use an assumption of spatial homogeneity (stationarity), so that 3D Fourier space coefficients could be computed independently, however one would need to take care of the statistical anisotropy in the LOS direction, therefore all statistics in Fourier space would depend on $|k_\perp|$ and $k_\parallel$. Taking into account transverse correlations would thus further reduce the covariance of the flux conditionally to the DM fields, in other words reduce the noise in the predicted flux field. Among future prospects, we plan to extend this work to predict the flux clustering for other surveys such as the *Dark Energy Spectroscopic Instrument* (DESI; DESI Collaboration et al.

2016), the *William Herschel Telescope Enhanced Area Velocity Explorer* (WEAVE-QSO; Pieri et al. 2016) or *Subaru Prime Focus Spectrograph* (PFS; Takada et al. 2014). They will open new vistas on the high redshift intergalactic medium probed by the Ly α forest. It would be then interesting to estimate the level of performance of LyMAS2 when the transmitted flux has a higher resolution than BOSS spectra, which might require reducing the DM smoothing. Finally, we also intend to use Machine Learning in the process to see whether we can still improve the predicted flux statistics (Chopitan et al., in preparation).

ACKNOWLEDGEMENTS

We warmly thank the referee for an insightful review that considerably improved the quality of the original paper. This work was carried within the framework of the Horizon project (http://www.projet-horizon.fr). Most of the numerical modelling presented here was done on the Horizon cluster at IAP. This work was supported by the Programme National Cosmology et Galaxies (PNCG) of CNRS/INSU with INP and IN2P3, co-funded by CEA and CNES. DW acknowledges support of US National Foundation grant AST-2009735. We warmly thank T. Sousbie, B. Wandelt, O. Hahn, M. Buehlmann, and S. Rouberol for stimulating discussions. We also thank D. Munro for freely distributing his Yorick programming language (available at http://yorick.sourceforge.net/) that was used during the course of this work.

DATA AVAILABILITY

The data and numerical codes underlying this paper were produced by the authors. They will be shared on reasonable request to the corresponding author.

REFERENCES

Bautista J. E. et al., 2017, A&A, 603, A12
Bertone S., White S. D. M., 2006, MNRAS, 367, 247
Bi H., Davidsen A. F., 1997, ApJ, 479, 523
Blanton M. R. et al., 2017, AJ, 154, 28
Bolton J. S., Puchwein E., Sijacki D., Haehnelt M. G., Kim T.-S., Meiksin A., Regan J. A., Viel M., 2017, MNRAS, 464, 897
Buehlmann M., Hahn O., 2019, MNRAS, 487, 228
Busca N. G. et al., 2013, A&A, 552, A96
Cai Z. et al., 2016, ApJ, 833, 135
Caucci S., Colombi S., Pichon C., Rollinde E., Petitjean P., Sousbie T., 2008, MNRAS, 386, 211
Chabanier S., Bournaud F., Dubois Y., Palanque-Delabrouille N., Yèche C., Armengaud E., Peirani S., Beckmann R., 2020, MNRAS, 495, 1825
Colombi S., Chodorowski M. J., Teyssier R., 2007, MNRAS, 375, 348
Croft R. A. C., Weinberg D. H., Katz N., Hernquist L., 1998, ApJ, 495, 44
Croft R. A. C., Weinberg D. H., Pettini M., Hernquist L., Katz N., 1999, ApJ, 520, 1
DESI Collaboration et al., 2016, preprint (arXiv:1611.00036)
Dalton G. et al., 2016, in Evans C. J., Simard L., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI. SPIE, Bellingham, p. 99081G
Dalton G. et al., 2020, Proc. SPIE Conf. Ser. Vol. 11447, Ground-based and Airborne Instrumentation for Astronomy VIII. SPIE, Bellingham, p. 1144714
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
Dubois Y. et al., 2014, MNRAS, 444, 1453
du Mas des Bourboux H. et al., 2020, ApJ, 901, 153
Eisenstein D. J. et al., 2011, AJ, 142, 72
Faucher-Giguère C.-A., Prochaska J. X., Lidz A., Hernquist L., Zaldarriaga M., 2008, ApJ, 681, 831
Font-Ribera A. et al., 2012, J. Cosmol. Astropart. Phys., 2012, 059
Font-Ribera A. et al., 2013, J. Cosmol. Astropart. Phys., 2013, 018
Font-Ribera A. et al., 2014, J. Cosmol. Astropart. Phys., 2014, 027
Francis P. J., Hewett P. C., 1993, AJ, 105, 1633
Gnedin N. Y., Hui L., 1996, ApJ, 472, L73
Harrington P., Mustafa M., Dornfest M., Horowitz B., Lukić Z., 2022, ApJ, 929, 160
Hockney R. W., Eastwood J. W., 1988, Computer Simulation Using Particles. Bristol, Hilger
Horowitz B., Dornfest M., Lukić Z., Harrington P., 2021, preprint (arXiv:2106.12675)
Japelj J. et al., 2019, A&A, 632, A94
Komatsu E. et al., 2011, ApJS, 192, 18
Kraljic K. et al., 2022, preprint (arXiv:2201.02606)
Lee K.-G. et al., 2015, ApJ, 799, 196
Lee K.-G. et al., 2018, ApJS, 237, 31
Lochhaas C. et al., 2016, MNRAS, 461, 4353
Lynds R., 1971, ApJ, 164, L73
McQuinn M., 2009, ApJ, 704, L89
Monaghan J. J., Lattanzio J. C., 1985, A&A, 149, 135
Ozbek M., Croft R. A. C., Khandai N., 2016, MNRAS, 456, 3610
Palanque-Delabrouille N. et al., 2013, A&A, 559, A85
Peeples M. S., Weinberg D. H., Davé R., Fardal M. A., Katz N., 2010, MNRAS, 404, 1281
Peirani S., Weinberg D. H., Colombi S., Blaizot J., Dubois Y., Pichon C., 2014, ApJ, 784, 11 (P14)
Peirani S. et al., 2017, MNRAS, 472, 2153
Pichon C., Vergely J. L., Rollinde E., Colombi S., Petitjean P., 2001, MNRAS, 326, 597
Pieri M. M. et al., 2016, preprint (arXiv:1611.09388)
Ravoux C. et al., 2020, J. Cosmol. Astropart. Phys., 2020, 010
Sargent W. L. W., Young P. J., Boksenberg A., Tytler D., 1980, ApJS, 42, 41
Schaye J. et al., 2015, MNRAS, 446, 521
Shi D., Cai Z., Fan X., Zheng X., Huang Y.-H., Xu J., 2021, ApJ, 915, 32
Sinigaglia F., Kitaura F.-S., Balaguera-Antolínez A., Shimizu I., Nagamine K., Sánchez-Benavente M., Ata M., 2022, ApJ, 927, 230
Slosar A. et al., 2011, J. Cosmol. Astropart. Phys., 2011, 001
Slosar A. et al., 2013, J. Cosmol. Astropart. Phys., 2013, 026
Sorini D., Oñorbe J., Lukić Z., Hennawi J. F., 2016, ApJ, 827, 97
Springel V., 2005, MNRAS, 364, 1105
Takada M. et al., 2014, PASJ, 66, R1
Teyssier R., 2002, A&A, 385, 337
Tie S. S., Weinberg D. H., Martini P., Zhu W., Peirani S., Suarez T., Colombi S., 2019, MNRAS, 487, 5346
Viel M., Schaye J., Booth C. M., 2013, MNRAS, 429, 1734
Vogelsberger M. et al., 2014, MNRAS, 444, 1518
Weinberg D. H., Cole S., 1992, MNRAS, 259, 652
Weinberg D. H., Hernsquit L., Katz N., Croft R., Miralda-Escudé J., 1997, in Petitjean P., Charlot S., eds, IAP Colloq. 13, Structure and Evolution of the Intergalactic Medium from QSO Absorption Line System. Nouvelles Frontières, Paris, p. 133
Weinberg D. H., Katz N., Hernquist L., 1998, in Woodward C. E., Shull J. M., Thronson Harley A. J., eds, ASP Conf. Ser. Vol. 148, Origins. Astron. Soc. Pac., San Francisco, p. 21

APPENDIX A: GENERAL TRENDS

In Section 5.1, we have presented the predictions regarding the two-point correlation functions for eight different combinations of DM fields, with calibrations derived from HORIZON-NOAGN. To check the robustness of the results, the analysis of other similar hydrodynamical simulations is definitely required. To limit the computational time, we ran five additional hydrodynamical simulations with the same boxside and same physics than HORIZON-NOAGN but with

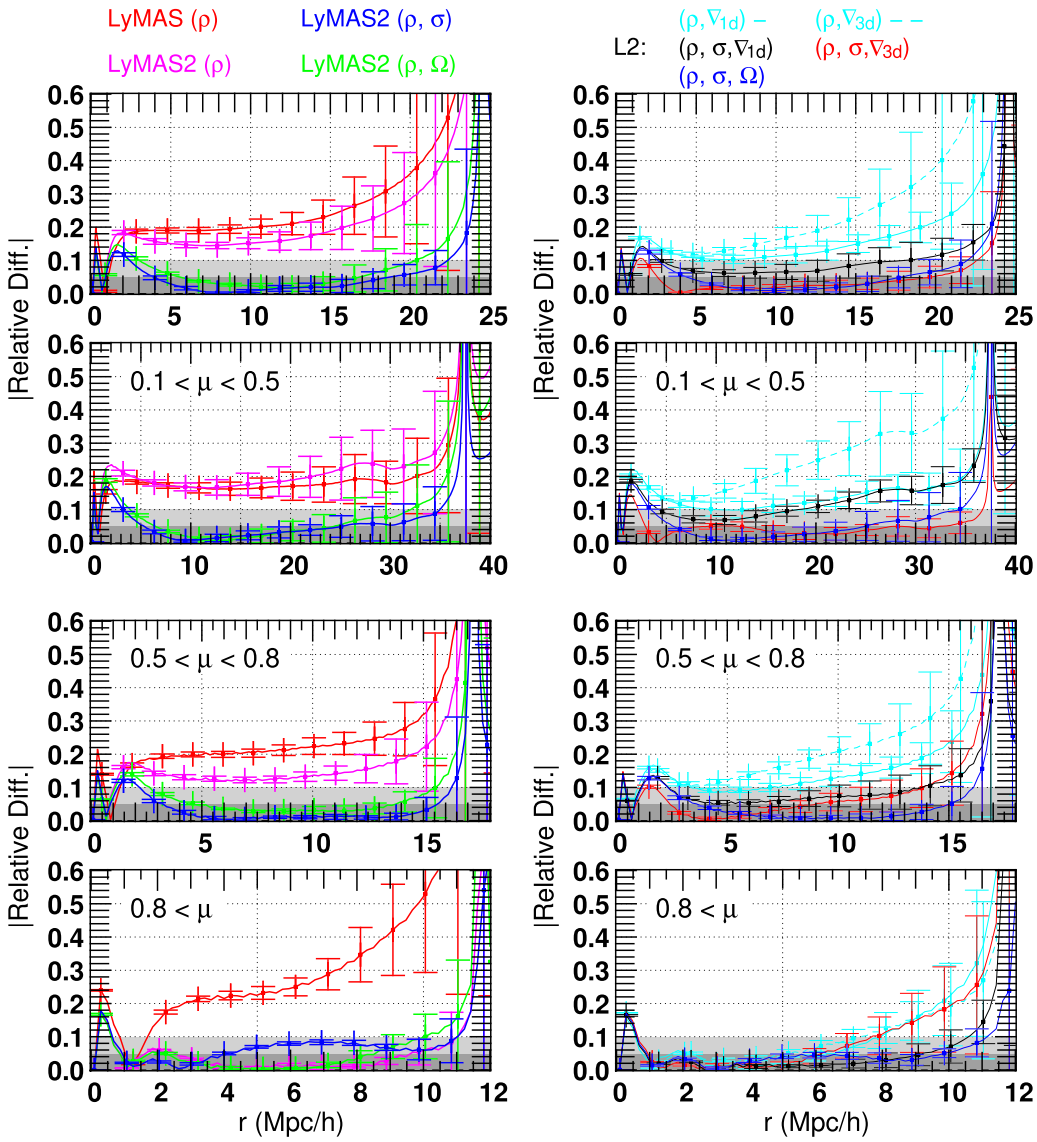

Figure A1. The evolution of the mean of the absolute relative difference in the two point correlation functions $(1/5)|\sum_{i=1}^{5}(\xi_i/\xi_{\mathrm{hydro},i}-1)|$ derived from five different hydrodynamical simulations at $z = 2.5$. All DM fields are smoothed to 0.5 Mpc h^{-1}. The error bars correspond to the dispersion. Comparison with results from Figs 7 and 8 suggest a very good agreement and therefore robust trends.

two times lower resolution (i.e. 512^3 DM particles instead of 1024^3 and a minimal cell size of $\Delta x = 2$ kpc instead of 1 kpc). The first simulation uses degraded HORIZON-NOAGN initial conditions, while the other ones have different initial phases. For each of the five new simulations, we generated the corresponding grids of transmitted flux, DM overdensity, and velocity fields and calibrations following the same methodology presented in Section 4. We consider here flux and DM fields sampled on grids of $512 \times 512 \times 1024$, namely 512×512 spectra of resolution 1024.

In the first step, we consider all DM fields smoothed at 0.5 Mpc h^{-1}. After checking first that the 'high' and 'low' resolution HORIZON-NOAGN simulations give consistent trends, we took an interest in the variations of then mean of the absolute relative difference $(1/5)|\sum_{i=1}^{5}(\xi_i/\xi_{\mathrm{hydro},i}-1)|$, where we compare the two-point correlation function of the hydro spectra $\xi_{\mathrm{hydro},i}$ from a given simulation 'i' to those derived from pseudo-spectra generated with LyMAS2 $\xi_{\mathrm{hydro},i}$. In Fig. A1, we summarize the results obtained with the original LyMAS and LyMAS2 considering the same DM field combinations than in Figs 7 and 8. The main conclusion is that we do find very similar trends than those obtained with HORIZON-NOAGN, which strongly suggest that our results are robust. In particular, the use of the velocity dispersion (σ) or the vorticity (Ω) lead to relative errors that are remarkably low, i.e. in general lower than 5 per cent even for the different ranges of angle. The plots also confirm that the 1D and 3D velocity divergence fields do no permit to reach the same level of accuracy.

In the next step, we present the trends obtained when the DM fields are smoothed to 0.3 or 1.0 Mpc h^{-1}. We only present in Fig. A2 the results for LyMAS, LyMAS2(ρ, σ) and LyMAS2(ρ, σ, Ω) to have a clear overview of the general trends. In P14, we found that a DM smoothing of 0.3 Mpc h^{-1} was an optimal value to reach the highest accuracy in the predictions. This is confirmed here since we get errors of $\geq$10 per cent compared to $\geq$20 and $\geq$30 per cent with values 0.5 and 1.0 Mpc h^{-1}, respectively. As expected, LyMAS2 permits to reduces such errors that are in general much lower than 10 per cent and most of the time lower than 5 per cent. It is also very promising

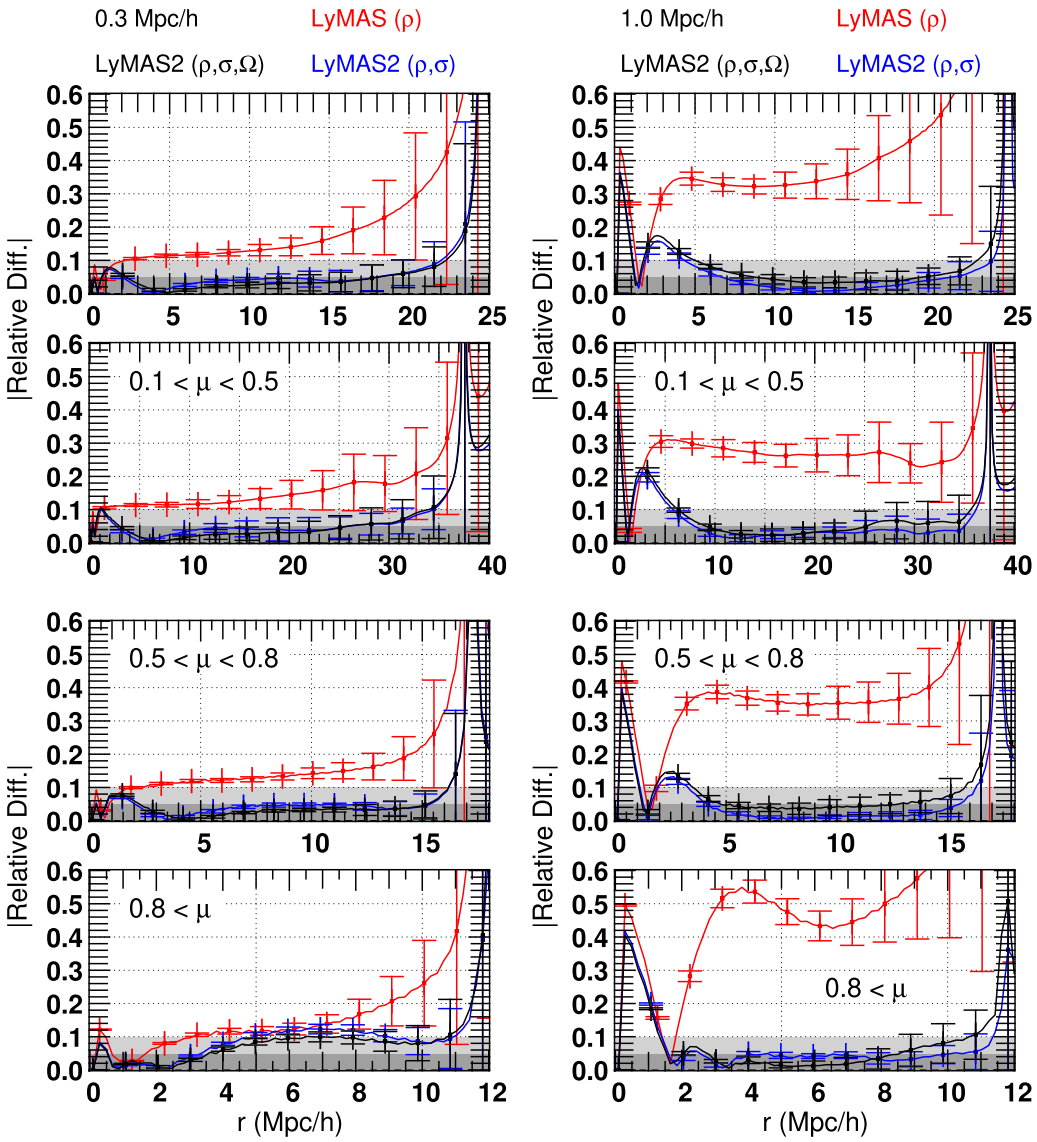

Figure A2. Same as Fig. A1 but for DM smoothing of 0.3 (left-hand column) or 1.0 Mpc h^{-1} (right-hand column). For the sake of clarity, we only show results for the original LyMAS (red curves), LyMAS2(ρ, σ) (blue curves), and LyMAS2(ρ, σ, Ω) (black curves). Note that even with a DM smoothing of 1.0 Mpc h^{-1}, LyMAS2 still permits to reach a high-level accuracy, even for high angles ($\mu > 0.8$).

that LyMAS2 applied to DM fields smoothed at 1 Mpc h^{-1} gives such accurate predictions even for high values of μ. This is definitely not the case with the original LyMAS leading to very high errors. Note also that due the smoothing scale, the predictions are less accurate for distance lower than 2 Mpc h^{-1} but acceptable for large-scale analysis.

APPENDIX B: DETERMINISTIC MAPPING

One commonly way to produce large mocks of Ly α forest, from Gaussian fields or DM distributions extracted from cosmological simulations, is to use a physically motivated deterministic relation that links the Ly α optical depth (or transmitted flux) to the DM overdensity. This is the case with the so-called Fluctuating Gunn–Peterson Approximation that has been extensively used in the literature. However, the FGPA is supposed to be more suitable for modelling high-resolution spectra and can be strongly limited when the DM density field is smoothed to a scale greater than 0.1 Mpc h^{-1} (see e.g. the analysis of Sorini et al. 2016) and confirmed by our results in Section 5.4. For this reason, we have derived in P14 an 'optimal' deterministic relations by matching the corresponding cumulative distributions of the smoothed transmitted Flux F_s and DM overdensity ρ_s as

$$\int_0^{F_s} P(F_s')\mathrm{d}F_s' = \int_{\rho_s}^{\infty} P(\rho_s')\mathrm{d}\rho_s' ,$$

where $P(F_s)$ and $P(\rho_s)$ are the one-point PDFs of the flux and DM overdensity measure from the simulation. One advantage of choosing such deterministic relation is to recover by construction the PDF of the hydro flux. However, the two-point correlation function of pseudo-spectra generated with this approach is still highly overestimated (see e.g. figs 10 and 19 in P14).

In this section, we consider another choices of deterministic relations. In particular, Tie et al. (2019) have used the conditional probability $P(F|1+\delta)$ of the transmitted flux on the DM overdensity

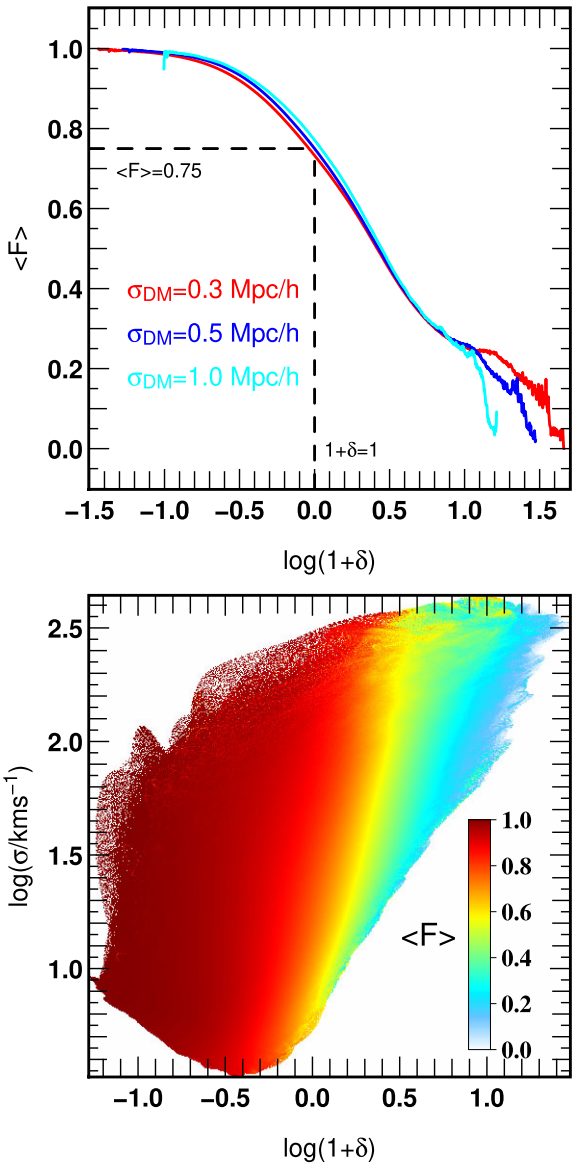

Figure B1. Examples of 1D and 2D deterministic relations derived from the HORIZON-NOAGN simulation at $z = 2.5$. Top panel: the mean flux $\langle F \rangle$ only depends on the DM overdensity $1 + \delta$. Three different DM smoothing have been considered. For example, the mean flux has a value of 0.75 for an overdensity of 1 (DM smoothing = 0.5 Mpc h^{-1}). Bottom panel: scatter plot showing the mean flux $\langle F \rangle$ with respect to the DM overdensity and velocity dispersion field (σ). In this case, $\langle F \rangle$ can have a wide range of values for $1 + \delta = 1$ (see also Fig. B2).

to get the conditional mean flux:

$$\overline{F}(1+\delta) = \int F.P(F|1+\delta)\mathrm{d}F\,. \tag{B1}$$

It can be analytically demonstrated that the two-point correlation function of pseudo-spectra obtained from such a deterministic mapping is the same that the one obtained with the first version of LyMAS (Tie et al. 2019). Since this deterministic relation can be easily extended to several DM fields, our aim is to investigate whether the inclusion of different DM velocity fields in such deterministic mappings may improve the trends or not. As an illustration, Fig. B1

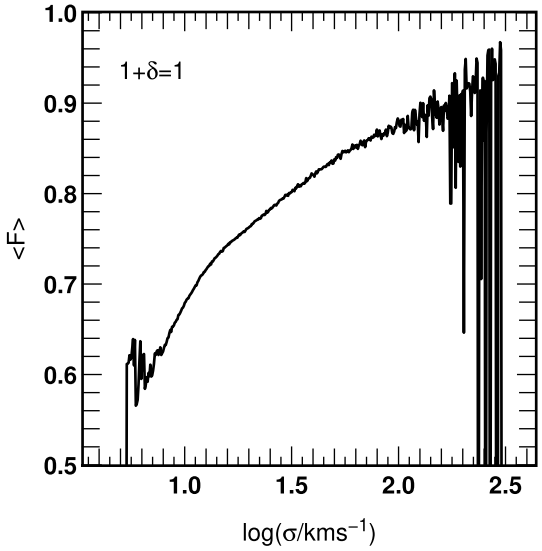

Figure B2. The evolution of the mean flux as respect to the velocity dispersion (σ) and for a given overdensity (i.e. $1 + \delta = 1$). The DM fields are smoothed at 0.5 Mpc h^{-1}. This variation is directly derived from the 2D deterministic sampling presented in Fig. B1. Compared to the 1D deterministic relation, a wide range of values of $\langle F \rangle$ is obtained and should refine the predictions.

shows examples of 1D-deterministic relations constructed for different smoothing of the DM overdensity field and and one example of 2D-deterministic relation $\overline{F}(1 + \delta, \sigma)$ using both the DM overdensity and the velocity dispersion fields (smoothed at 0.5 Mpc h^{-1}). All relations are derived from the HORIZON-NOAGN simulation at $z = 2.5$ (in redshift space). In principle, the 2D-deterministic relation is supposed to refine the results as respect to the 1D deterministic one. Indeed, let us take, for instance, a DM overdensity of $1 + \delta = 1$. This leads to an unique mean flux of $\langle F \rangle = 0.75$ from the 1D deterministic relation (using a DM smoothing of 0.5 Mpc h^{-1}). The 2D-deterministic relation provides, however, a wide range of possible values of $\langle F \rangle$ depending this time on the velocity dispersion (see Fig. B2).

We have then produced grids of pseudo-spectra from DM fields extracted from the HORIZON-NOAGN simulation (see Section 4), smoothed at 0.5 Mpc h^{-1} and using three different deterministic relations. The first one considers the DM overdensity field only (one-field), the second one both overdensity and velocity dispersion fields (two-fields), while the last one associates the DM density field to the velocity dispersion and vorticity fields (three-fields). To estimate the mean value of the flux from a given value of ρ or a given set of (ρ, σ) or (ρ, σ, Ω), we use, respectively, interpolations, bilinear interpolations, and trilinear interpolations, depending on the number of input DM fields. First, Fig. B3 shows the 1D power spectrum and PDF of pseudo-spectra (without iteration) for the two-fields case. We notice that the one-point PDF is in general not well recovered. The predictions of the 1D-P_k are also not as accurate than LyMAS and things especially for the two-fields and three-fields cases. Indeed, the power spectra at small scales are considerably overestimated. Although a full iteration can improved these trends, the predicted two-point correlation functions, shown in Fig. B4, present errors that are generally quite high especially when an angle μ is considered. For instance, the errors are much higher than 10 per cent when $\mu > 0.8$.

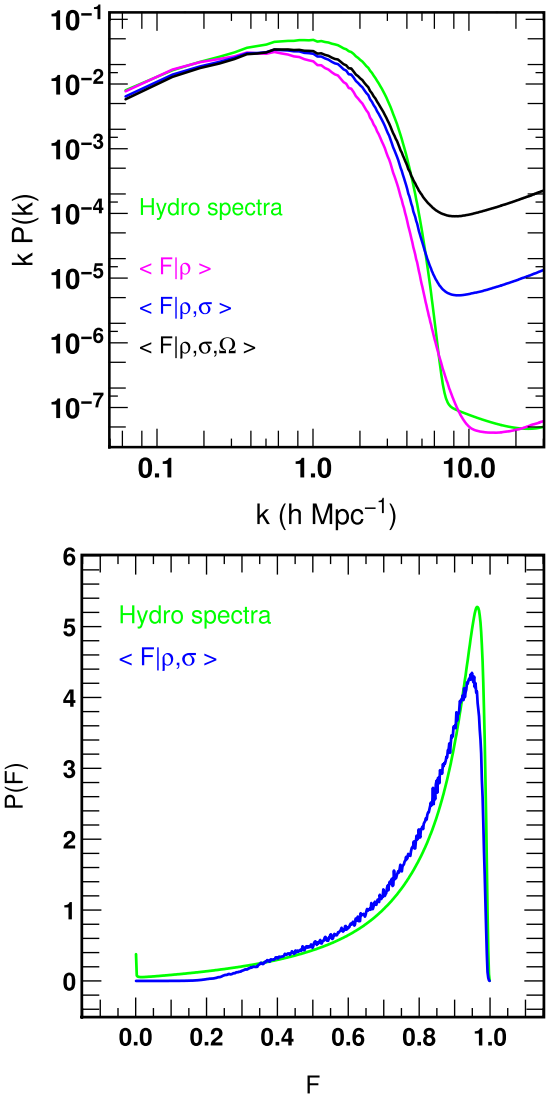

Figure B3. Top panel: the 1D power spectrum of pseudo-spectra generate from HORIZON-NOAGN DM fields (smoothed at 0.5 Mpc h^{-1}; $z = 2.5$) and using deterministic relations described in the text. The discrepancies with the hydro 1D-P_k are quite pronounced especially at small scales. Bottom panel: an example of PDF of pseudo-spectra compare to hydro spectra, showing again a noticeable disagreement. All results are presented without a full iteration in the scheme (flux 1D-P_k and PDF rescaling).

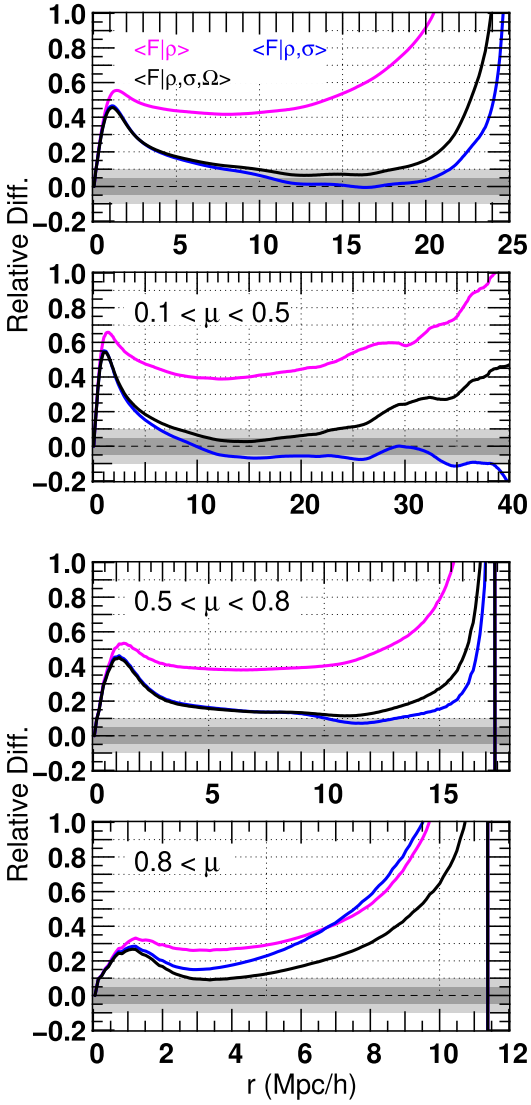

Figure B4. The relative difference of the two-point correlation functions of pseudo-spectra produced with deterministic sampling (and a full iteration in the scheme). It appears clearly that the different deterministic sampling do no reach the level of accuracy of LyMAS2 especially for high angles where the error are quite high (> 10 per cent). The light and dark grey shade represent error lower that (absolute) 10 and 5 per cent, respectively.

Also, one main drawback of this approach when creating large mocks (>1 Gpc h^{-1}) is to have enough statistics from the hydrodynamics simulations to cover most of the parameter space of the large cosmological simulation. This should not be an issue for the one-field case since simple interpolations and extrapolations can be done (e.g. from Fig. B1, $\langle F \rangle \sim 1$ and $\langle F \rangle \sim 0$ for $1 + \delta < -1.5$ and $1 + \delta > 1.5$, respectively). For the two-fields and three-fields cases, efficient interpolations and extrapolations could be obviously much more complicated to realized.

APPENDIX C: ADAPTIVE SMOOTHING

In this section we provide some detail on how the DM density field, velocity field, and velocity dispersion are interpolated adaptively on a mesh from the output of a cosmological DM *N*-body simulation, before further treatment by LyMAS2, in particular additional smoothing with a Gaussian window.

Note that, while the notations below assume standard configuration space, the calculation naturally extend to redshift space, by just modifying particle coordinates accordingly with the local peculiar velocity component contribution. In LyMAS2, we use the infinitely remote observer approximation by just accounting for redshift distortion along the z-axis.

For a smooth phase-space distribution function $f(\boldsymbol{x}, \boldsymbol{v})$, the projected density is given by

$$\rho_{\rm E}(\boldsymbol{x}) = \int {\rm d}^3 v\ f(\boldsymbol{x}, \boldsymbol{v}), \tag{C1}$$

the Eulerian mean velocity field by

$$\boldsymbol{v}_{\rm E}(\boldsymbol{x}) = \langle \boldsymbol{v} \rangle_{\rm E} = \frac{1}{\rho_{\rm E}(\boldsymbol{x})} \int {\rm d}^3 v\ \boldsymbol{v}\ f(\boldsymbol{x}, \boldsymbol{v}), \tag{C2}$$

and the local mean square velocity reads

$$\langle v^2 \rangle_{\rm E}(\boldsymbol{x}) = \frac{1}{\rho_{\rm E}(\boldsymbol{x})} \int {\rm d}^3 v\ v^2\ f(\boldsymbol{x}, \boldsymbol{v}). \tag{C3}$$

Obviously, the last two equations stand for points of space where $\rho(\boldsymbol{x}) > 0$. From equation (C3), we can derive the local velocity dispersion

$$\sigma^2_{v,{\rm E}} \equiv \langle v^2 \rangle_{\rm E} - v^2_{\rm E}. \tag{C4}$$

We notice that the local velocity field can be considered as a statistical average, this is why we used the $\langle \cdots \rangle_{\rm E}$ notation above,

$$\langle \boldsymbol{v} \rangle_{\rm E}(\boldsymbol{x}) = \int {\rm d}^3 v\ \boldsymbol{v}\ f_{\rm E}(\boldsymbol{x}, \boldsymbol{v}), \tag{C5}$$

with the local density probability

$$f_{\rm E}(\boldsymbol{x}, \boldsymbol{v}) \equiv \frac{1}{\rho_{\rm E}(\boldsymbol{x})} f(\boldsymbol{x}, \boldsymbol{v}). \tag{C6}$$

In this probabilistic approach, the mean square velocity is given by equation (C3), since

$$\langle v^2 \rangle_{\rm E}(\boldsymbol{x}) = \int {\rm d}^3 v\ v^2\ f_{\rm E}(\boldsymbol{x}, \boldsymbol{v}). \tag{C7}$$

and, likewise, its local variance by equation (C4).

What we have actually access to is not a smooth distribution function, unfortunately, but a distribution of N simulation particles of individual masses m_i, positions $\boldsymbol{x}_i$ and velocities $\boldsymbol{v}_i$. This means that the phase-space distribution function has the following form:

$$f(\boldsymbol{x}, \boldsymbol{v}) = \sum_i m_i\ \delta_{\rm D}(\boldsymbol{v} - \boldsymbol{v}_i)\, \delta_{\rm D}(\boldsymbol{x} - \boldsymbol{x}_i), \tag{C8}$$

where $\delta_{\rm D}$ is the Dirac distribution function. From equation (C8), one can compute the projected Eulerian density

$$\rho_{\rm E}(\boldsymbol{x}) = \sum_i m_i\ \delta_{\rm D}(\boldsymbol{x} - \boldsymbol{x}_i), \tag{C9}$$

but the Eulerian velocity field $\boldsymbol{v}_{\rm E}(\boldsymbol{x})$ and mean square velocity are ill defined.

However, the underlying distribution of true DM particles is much smoother than its crude numerical representation in terms of macroparticles of the N-body simulation. While, strictly speaking, the phase-space density is still of the form (C8) at the microscopic level, it can be considered as a smooth function at the macroscopic level, at least in terms of probability density.

In order to recover a good approximation of the continuum, Colombi et al. (2007) proposed a locally adaptive smoothing algorithm, SmoothDens, inspired from smooth particle hydrodynamics (hereafter SPH), using, to compute various fields, an interpolation window $F_{\boldsymbol{x}}$, of which the shape parameters, in particular the typical size $\ell(\boldsymbol{x})$, depend on position $\boldsymbol{x}$. This window function is normalized to unity, i.e. $\int {\rm d}^3 x'\ F_{\boldsymbol{x}}(\boldsymbol{x}') = 1$.

In principle, for a given function $h(\boldsymbol{x}, \boldsymbol{v})$, the smoothed counterpart is given by

$$[F_{\boldsymbol{x}} \times h](\boldsymbol{x}, \boldsymbol{v}) = \int {\rm d}^3 x'\ F_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}')\, h(\boldsymbol{x}', \boldsymbol{v}). \tag{C10}$$

Setting $h = \rho_{\rm E}$, after simple algebraic calculations exploiting the properties of the Dirac distribution function, we obtain the simple expression for the adaptively smoothed density:

$$\rho_{\rm F}(\boldsymbol{x}) = \sum_i m_i\ F_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_i). \tag{C11}$$

From there, we can formally define the analogous of equation (C6) but with adaptive smoothing performed in the spatial position,

$$f_{\rm F}(\boldsymbol{x}, \boldsymbol{v}) \equiv \frac{1}{\rho_{\rm F}(\boldsymbol{x})} [F_{\boldsymbol{x}} \cdot f](\boldsymbol{x}, \boldsymbol{v}), \tag{C12}$$

from which one can derive estimates in the mean field limit of velocity related quantities:

$$\boldsymbol{v}_{\rm F}(\boldsymbol{x}) = \frac{\sum_i m_i\ \boldsymbol{v}_i\ F_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_i)}{\sum_i m_i\ F_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_i)}, \tag{C13}$$

$$\langle v^2 \rangle_{\rm G} = \frac{\sum_i m_i\ v_i^2\ F_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_i)}{\sum_i m_i\ F_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_i)}. \tag{C14}$$

In the algorithm SmoothDens, the adaptive procedure is used to compute various fields for a particular set of positions $\boldsymbol{x} = \boldsymbol{x}_j$ on a cubical grid of size $n_{\rm g}$. Function F is a compact (Monaghan & Lattanzio 1985) spline of size $\ell(\boldsymbol{x})$, with $\ell(\boldsymbol{x})$ being the distance of the $N^{\rm th}_{\rm SPH}$ closest simulation particle to position $\boldsymbol{x}$. The value of $N_{\rm SPH}$ we adopt here is $N_{\rm SPH} = 32$. As an additional recipe, adaptive smoothing is locally replaced with nearest grid point (NGP) interpolation (Hockney & Eastwood 1988) when $\ell(\boldsymbol{x})$ is smaller or of the order of the grid cell size $L/n_{\rm g}$, where L is the simulation box size. Also, a local weight is given to each particle i so that at the end, its total contribution sums up to the particle mass m_i. Note that, due to the finite extension of the spline function, some particles belonging to dense clusters or close to dense clusters may not contribute at all. In the latest implementation of the algorithm, SmoothDens5, which we use in LyMAS2, an option allows one to affect these particles to the grid with NGP interpolation in order to conserve total mass. The effect of not doing so is however generally small.

The outcome of SmoothDens mainly depends on two parameters, the resolution $n_{\rm g}$ of the grid and the value used for the number of neighbours, $N_{\rm SPH}$. Changing both these parameters can have drastic impact on the results, especially the local velocity dispersion and the velocity derivatives estimates. Additional Gaussian smoothing performed in LyMAS2 is however expected to reduce considerably the dependence on these two parameters, provided that the smoothing scale $R_{\rm G}$ associated with the smoothing window is large enough compared to $L/n_{\rm g}$. Yet one has to bear in mind that the influence of $N_{\rm SPH}$ cannot be negligible in underdense regions as long as it can influence scales larger than $R_{\rm G}$, which is unfortunately very likely. Despite these non-trivial issues, the reason why LyMAS2 still works so accurately is that it is calibrated relying on probability distributions mappings, which naturally corrects for intrinsic biases introduced by local adaptive smoothing.

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