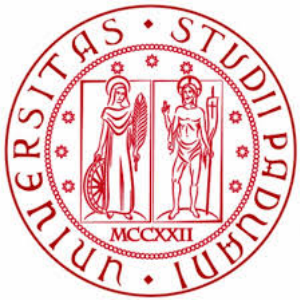

Università degli Studi di Padova

Dipartimento di Fisica e Astronomia

Master Degree in Physics

Final Dissertation

Mapping Large Scale Structure tracers onto the dark matter cosmic web

Candidate: Francesco Sinigaglia
Candidate number: 1207013

Thesis supervisor:
Prof. Michele Liguori
Thesis co-supervisors:
Prof. Francisco-Shu Kitaura Joyanes
Dr. Andrés Balaguera-Antolínez

Academic Year 2019/20

Alle mie nonne, Carla e Maria
Ai miei genitori

ABSTRACT

Dark matter, galaxies and the intergalactic medium form on large scales a filamentary pattern, called cosmic web. In this context, collapsed structures, such as galaxies, galaxy clusters and dark matter halos, reside preferentially in overdense regions of the matter density field and are hence, regarded as tracers of the underlying dark matter distribution.
In this work we apply the BAM method to obtain the mapping relation between the dark matter density field and the Large Scale Structure tracers, and ultimately to sample accurate mock catalogs. In the first part, we extract the halo bias relation in a stochastic sense from dark matter only N-body cosmological simulations, the MINERVA suite. We find that a proper re-definition of the cosmic web, dubbed I-web and based on the invariants of the tidal field tensor, allows to sample halo mock catalogs with unprecedented accuracy in the 2- and 3-point statistics. In the second part of the work BAM is applied to a zoom-in hydrodynamical simulation, to learn how to map ionized gas density, neutral hydrogen (HI) number density, temperature and optical depths onto the dark matter field. We explore the possibility of exploiting the mapping between various baryonic tracer fields and the baryonic cosmic web classification, finding that the gas cosmic web carries more information on the spatial distribution of HI than the dark matter cosmic web. The final goal of these studies consists of sampling a great number of accurate mock catalogs of halos, IGM properties and Lyman-α forest fluxes, to be delivered to forthcoming surveys such as DESI, EUCLID, and J-PAS.

CONTENTS

INTRODUCTION

Large observational sky-mapping campaigns show that galaxies, dark matter and the intergalactic medium form on large scales at low-redshift a filamentary pattern connected through knots, and leaving large void regions, conforming the so-called *cosmic web*. Current models of structure formation must be able to account for such observed patterns and show that gravitationally collapsed clumps of dark matter, called *dark matter halos*, form in a complex dependence on the dark matter density field. In the current view of this field, smaller structures form first and then cluster together, giving rise to larger structures, in a so-called *hierarchical clustering scenario*. Therefore, cosmic structures we observe at low-redshift in the Universe are thought to form as a result of the evolution and growth across cosmic time from initial (post-inflationary) density perturbations, as a result of gravitational instability. Galaxies and galaxy clusters, are in turn thought to form within dark matter halos involving additional complex baryonic physics, such as star formation, AGN feedback, Supernova feedback, and cooling among others.

The acquisition of this knowledge is mainly based on a combined analysis of observations and expensive hydro-dynamical simulations. However, to make cosmological analysis larger volumes than the ones obtained in these simulations need to be addressed. To this end, one needs to resort to more efficient modelling techniques. Therefore, a huge theoretical effort is done employing a combination of analytical calculations, numerical simulations and semi-analytical models, which are far less expensive.

Thanks to the large amount of data that current surveys, such as, e.g., eBOSS (Dawson et al., 2016) and DES (The Dark Energy Survey Collaboration, 2005), and forthcoming observational campaigns, such as, e.g., Euclid (Amendola et al., 2018) and DESI (Levi et al., 2013), will provide to the scientific community, we are today in the privileged position of being able to map the Universe up to high redshifts, doing precision Cosmology, estimating cosmological parameters and hence, setting constrains to cosmological models of our Universe.

Besides improved maps of galaxies and galaxy clusters, future surveys will also provide unprecedented maps of the intergalactic medium. This is possible thanks to the so-called *Lyman-α* forest, which consists in a thick ensemble of absorption lines characterizing spectra of distant quasars, located at $z > 2$, generated by the Lyman-α transition of neutral hydrogen. The Lyman-α forest turns out to be an excellent tracer of the IGM at $z \sim 2 - 3$, where normal galaxies become too faint to be observed. DESI will indeed

measure angular and redshift coordinates of Luminous Red Galaxies up to $z \sim 1$ and Emission Line Galaxies up to $z \sim 1.6$, while it will be able to detect quasars up to $z \sim 3.5$ and is expected to extend the eBOSS database of quasars spectra by almost an order of magnitude. Such quasar spectra will not only allow us to infer information on the quasars themselves, but also to reconstruct the column density of neutral hydrogen along the line-of-sight.

Another valuable source of IGM data is represented by large radio-surveys, which will trace for instance the distribution of HI in emission, by observing the so-called 21-cm emission line. Even though this goes far beyond the scope of this work, a number of studies (see e.g. Carilli et al., 2004; Kooistra et al., 2017, and references therein) showed that future radio-telescopes arrays, such as the Square Kilometre Array (see e.g. Weltman et al., 2020), will be able to detect the 21-cm emission line up to redshift $z \sim 6$ and beyond and probe the epoch of Reionization.

All the above mentioned surveys require covariance matrices and estimates of errors on cosmological observables. The standard approach to perform such estimations has become the building of large sets of mock catalogs. The first attempt for producing these mock catalogs consisted in running full N-body cosmological simulations. However, given the large computational cost and time required by such a procedure, several methods have been presented in the literature (see chapter 4), to speed up the mock generation. The most popular application is the building of dark matter halo mock catalogs. These can then be used for creating galaxy mock catalogs by populating halos with galaxies, according to some Halo Occupation Distribution (HOD) or Sub-Halo Abundance Matching (SHAM) prescription.

In this work we present the application of the `BAM` method to the creation of halo mock catalogs and mock catalogs of properties of the IGM, including Lyman-α forest fluxes. `BAM` is based on the idea of mapping a certain class of Large Scale Structure tracers onto the dark matter field by means of a statistical stochastic bias relation. Along this work I will present how the method is pushed to high accuracy, allowing to produce mock catalogs which are indistinguishable from the reference simulation in terms of various statistical measures. To this end, a central role is played by the dark matter cosmic web, which, in different possible definitions, is regarded to carry important information on the spatial distribution on large scales of the tracers of the dark matter field.

Therefore, the purpose of this work is the following: learn how to map different classes of Large-Scale Structure tracers, namely dark matter halos and constituents of the intergalactic medium, onto the dark matter field, in the context of methods for mock-production. For this purpose, we show that an accurate modelling of the underlying dark matter cosmic web is fundamental and allows us to retrieve important physical insights.

This work is organized as follows:

- chapter 1 presents the main basic concepts of Cosmology and highlights why they are linked to the matter of this thesis;
- chapter 2 introduces some basic knowledge about the halo bias, the IGM and the Lyman-α forest, which represent the topics which this work is focused on;
- in chapter 3 I present in details the topic of mock catalogs production, the details of the `BAM` code, how it is applied for building dark matter halo mocks and how it is extended to deal with continuous properties of the IGM;
- in chapter 4 I explain how `BAM` is applied to the study of halo bias and report on the results we included in a recently submitted paper (Kitaura et al., 2020);
- chapter 5 presents how `BAM` is applied to a hydrodynamical simulation, for the purpose of building mocks catalogs of some properties of the IGM and Lyman-α forest fluxes, in the view of DESI.

I will often refer to dark matter as DM and to the intergalactic medium as IGM throughout the work.

This manuscript has been typeset using LaTeX.
The work has been carried out mainly at Instituto de Astrofísica de Canarias (Tenerife, Spain), under the supervision of Prof. Francisco-Shu Kitaura and Dr. Andrés Balaguera-Antolínez.

1 BASIC CONCEPTS OF COSMOLOGY

In this chapter we provide an overview of the main basic concepts of modern Cosmology. We will present the main components of our Universe, the Cosmological Principle, the Robertson-Walker metric and the Friedmann equations, the Hubble expansion, the main cosmological models, the cosmological parameters and the standard ΛCDM model. Finally, we will mention how the structure formation process occurs, the concept of *hierarchical clustering* of cosmic structures and the 2- and 3-point summary statistics, in configuration and Fourier space. In particular, we will define the correlation function, the power spectrum and the bispectrum and will introduce which are the estimators which are commonly adopted to estimate such objects. The study of summary statistics in Fourier space has a prominent role in this work in the assessment of the accuracy of our predictions, since we rely on power spectrum and bispectrum to perform quantitative analyses of the results of our numerical experiments.

1 WHAT IS OUR UNIVERSE MADE OF?

To begin with, one could wonder which are the physical entities which constitute our Universe. According to the recent experimental results by the Planck mission and to the standard ΛCDM model (discussed later on in this chapter), we can subdivide the energetic budget of the present-day Universe as follows:

- baryonic matter, $\sim 4.7\%$, consisting mainly in light elements such as H, He and Li, which are thought to have formed during the so-called *Big-Bang Nucleosynthesis* and whose observed abundances are confirmed by theoretical predictions;
- dark matter, $\sim 27\%$, a matter component which is regarded to be made of non-interacting non-baryonic massive particles. The existence of dark matter has been supported by many observational evidences, such as, e.g. the rotation curves of galaxies and the lensing effect produced by gravity clusters, and turns out to be fundamental to explain the process of structure formation. Huge theoretical and experimental efforts have tried to unveil the nature of dark matter, but it has not been directly detected yet;
- dark energy, $\sim 68.3\%$, is by far the most abundant but also the least understood component of our Universe. Dark energy has been theo-

rized to explain the accelerated phase of expansion the late Universe is experiencing, but very little is know about it;

- radiation and neutrinos, in a negligible fraction.

However, the relative energetic importance of such components across cosmic time had not been always as it is today. On the contrary, different cosmic epochs have been energetically dominated by different components. In order to address quantitatively how our Universe has evolved, from its birth to the present day, we will introduce in the following sections the main mathematical tools commonly used to described the thermal history of the Universe.

2 THE COSMOLOGICAL PRINCIPLE

The Cosmological Principle is the foundational hypothesis of Cosmology and states that, on sufficiently large scales, the Universe is homogeneous and isotropic. This means that, on such scales, no matter in which direction nor at which point one observes the Universe: its properties will always look identical.

The Cosmological Principle is supported by strong observational evidences. The Large Scale Structure, i.e. the spatial distribution of matter at scales $l \gtrsim 100\,\mathrm{Mpc}$, supports the hypothesis of homogeneity. At smaller scales, the non-linear evolution of cosmic perturbations favours the formation of collapsed structures, which tend to organize themselves in an inhomogeneous pattern called *cosmic web* (see chapter 2). However, at the mentioned large scales, the Universe appears statistically homogeneous.

The isotropy hypothesis is instead supported by the *Cosmic Microwave Background* (CMB), which is the object in nature with the energy spectral distribution which resembles more closely the one of a perfect Black Body, and is found to be characterized by very tiny temperature anisotropies, of the order of $\Delta T/T \sim 10^{-5}$.

Of course, our knowledge about the Universe is limited to the cosmic volume around us that we can directly investigate with observations. In principle, one cannot be completely sure that the Cosmological Principle holds everywhere in the Universe. However, the common approach is to assume its validity and exploit the symmetries it introduces.

3 ROBERTSON–WALKER METRIC AND FRIEDMANN EQUATIONS

Let us assume the validity of the Cosmological Principle and adopt a set of co-moving radial coordinates $\vec{r}$. Co-moving coordinates are coordinates which moves together with the cosmological fluid. A freely-moving fluid element is at rest in the co-moving coordinate system, and all the observers who are at rest with the local freely-moving fluid element are called fundamental observers.

Let then $\vec{x}$ indicate physical coordinates and $a(t)$ be the *scale factor*. The scale factor $a(t)$ measures basically how physical separations between cosmic objects grow with time and relates co-moving and physical coordinates through the relation $\vec{r} = a(t)\,\vec{x}$.

The metric which is usually adopted to describe the geometrical properties of the Universe is the *Robertson-Walker metric*, defined as

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{1}{c^2}\frac{a(t)^2}{a_0^2}\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi)\right] \quad , \tag{1}$$

where r is the radial coordinate, θ and ϕ are the two (co-moving) Eulerian angles, a_0 is the present-day scale factor and the $(-,+,+,+)$ signature has been adopted. Such metric describes effectively an expanding space-time characterized by maximally-symmetric space-like hypersurfaces of curvature k.

The so-called *Friedmann equations* regulate the dynamics and the evolution of the energetic budget of the Universe and are given by the following relations:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} \tag{2}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) \quad , \tag{3}$$

where ρ is the matter density ρ_{m}, G is the gravitational constant, k is the spatial curvature and p is the matter pressure. In this case, we are assuming that the Universe is energetically and dynamically dominated by matter, which holds true for redshift $0.5 \lesssim z \lesssim 3000$. At such redshifts, the Universe is said to be *matter-dominated*. When instead one considers also Dark Energy, an additional term arises, which can be accounted for as a separate term, or can be re-absorbed in ρ. In the latter case, ρ indicates no longer matter density, but more in general the energy density of the component of the Universe which is dominant at the considered cosmic time.

The Friedmann equations can be obtained, with some assumptions, in the Newtonian limit but are consistently derived from the Einstein's field equations of General Relativity. The first equation expresses a formulation

of energy conservation in an expanding spacetime, while the second one is in fact a dynamical equation. As we will show later on in this chapter, an expanding Universe keeps the value k unchanged during all its evolution.

The Friedmann equations are two equations of 3 unknown variables: $\rho(t)$, $p(t)$ and $a(t)$. Therefore, to solve them one has to specify a third independent equation, an equation-of-state (EoS), which in many cases of physical interest is a linear relation $p = w\,\rho$ between density and pressure. In general, the Universe is approximated as being filled by a cosmological fluid made of different energetic components, each one with its EoS. The EoS is specified by the choice of w, where:

- $w = 0$ for matter: pressure-less dust EoS, valid for a fluid of non-relativistc particles;
- $w = 1/3$ for radiation: valid for a fluid of ultra-relativistic non-degenerate particles in thermal equilibrium, which exert a non-negligible pressure;
- $w = -1$ for the cosmological constant Λ, i.e. a possible EoS for dark energy.

By performing all the calculations, assuming $k = 0$ and that the Universe is dominated by a single perfect fluid with proper EoS, one finally finds the following solutions of the Friedmann equations:

$$\begin{aligned}
\text{Radiation-dominated Universe:} \quad & a(t) \propto t^{1/2}\,, \quad \rho(a) \simeq \rho_{\rm rad}(a) \propto a^{-4} \quad , \\
\text{Matter-dominated Universe:} \quad & a(t) \propto t^{2/3}\,, \quad \rho(a) \simeq \rho_{\rm m}(a) \propto a^{-3} \quad , \\
\Lambda\text{-dominated Universe:} \quad & a(t) \propto e^{H\,t}\,, \quad \rho \simeq \rho_\Lambda = \text{const.}
\end{aligned}$$

Later on in this chapter we will see that such solutions represent different evolutionary phases of the Universe, as predicted by the ΛCDM model.

4 THE EXPANDING UNIVERSE

Before proceeding with the presentation of the main cosmological models, a brief note on the expansion of Universe is needed. Many observational facts indicate that our Universe is expanding today. The meaning of expansion is that distant galaxies and objects are getting further apart. The expansion of the Universe becomes significant and observationally evident when two objects recede due to the fact that their motion is not being governed by their mutual gravitational force. As we will see in the following sections, the Universe does not expand in the same way across all the cosmic time. On the contrary, the fact that it is expanding today does not guarantee it will expand forever, but it may also stop its expansion and even re-contract,

according to what some cosmological models predict. Current experimental data indicate that the late Universe has entered a phase of accelerated expansion.

5 COSMOLOGICAL MODELS

In this section we examine the cosmological models without the cosmological constant Λ, which will be included in section 6. Moreover, we consider here only matter-dominated models, which well represent the evolution of the Universe after the very early evolutionary phases, i.e. after the matter-radiation Equivalence, which took place at $z \sim 3000$.

The cosmological models consist in possible solutions of the Friedmann equations and differ one another for the underlying geometry of the Universe. In particular, the spatial curvature k can assume the following values:

- $k < 0$, in the case of an open Universe;
- $k > 0$, in the case of a closed Universe;
- $k = 0$, in the case of a flat Universe.

Usually the scale factor is re-scaled, so that $k = 1$ if the Universe is open and $k = -1$ if it is closed.
Let us use this notation and assume $k = 0, \pm 1$. The resulting cosmological models are the following.

5.1 The Einstein-de Sitter Universe

This model corresponds to a spatially flat Universe ($k = 0$), i.e. the spacetime is a spatially-flat 4-dimensional manifold. In this case equation 2, the first Friedmann equation, loses its dependence on k and can be solved by simply separating variables, resulting in $a(t) \propto t^{2/3}$. This means that if $t \to 0$ then $a(t) \to 0$ and if $t \to \infty$ then $a(t) \to \infty$, but in a decelerated way.

5.2 The open model

This model of Universe is characterized by $k = -1$ and this makes both the right-hand terms of equations 2 positive. As a result, the left-hand term $\dot{a}/a$ never becomes vanishing nor negative. This fact means that the expansion is infinite and monotonic and hence $a(t) \to \pm ct$ when $t \to \infty$.

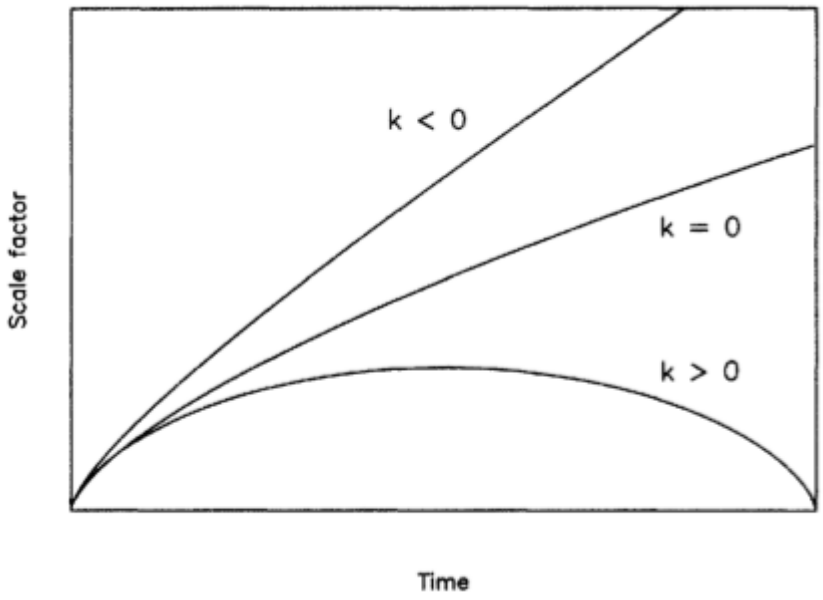

Figure 1: This plot shows the evolution of the scale factor for an Einstein-de Sitter Universe ($k = 0$), for an open Universe ($k < 0$) and for a closed Universe ($k > 0$). From Liddle (1998).

5.3 The closed model

In a closed Universe, at a certain $\tilde{t}$ the first derivative of the scale factor $\dot{a}(\tilde{t}) = 0$. As a result of this fact and of the second derivative $\ddot{a} < 0 \ \forall t$, the Universe expands until $t = \tilde{t}$, arrests its expansion at $\tilde{t}$ and then starts to contract. At this point, the contraction proceeds until the Universe collapses in one point and performs a so-called *Big Crash*.

Figure 1 shows the evolution of the scale factor as a function of time for the three discussed models. It turns out that all these models share the initial singularity, known as *Big Bang*, but evolve in different ways across cosmic time.

5.4 Cosmological parameters

In this section we introduce the cosmological parameters. We do not consider them all here, but focus only on some of them, which turn out to be useful for our purposes:

- the Hubble parameter $H(t)$. It is defined as $H(t) \equiv \dot{a}(t)/a(t)$ and can be interpreted as a measure of the expansion rate of the Universe at cosmic time t. The Hubble constant H_0 is the Hubble parameter evaluated at the present time t_0. The galaxies in the Local Universe ($z \ll 1$, where $z = 0$ is the redshift of the observer) had been experimentally found to follow a linear relation between their recession velocity

(velocity component along the line-of-sight) and their proper distance, with proportionality constant H_0 (Hubble, 1929). More precisely, this relation is known as *Hubble's law* and is usually expressed as

$$v = H_0\, d \quad , \tag{4}$$

where v is the recession velocity, H_0 the Hubble constant and d the proper distance from the origin of our system of coordinates;

- the deceleration parameter $q(t)$, defined as

$$q(t) \equiv -\frac{a(t)\,\ddot{a(t)}}{\dot{a}(t)^2} \quad . \tag{5}$$

It is related to the second derivative of the scale factor and it is named after the acceleration (or, strictly speaking, deceleration) of the cosmic expansion;

- the density parameter $\Omega_{\mathrm{m}}(t)$, whose name arises from its linear proportionality to the matter density of the Universe. It is defined as

$$\Omega_{\mathrm{m}}(t) \equiv \frac{8\pi\, G}{3\, H(t)^2}\, \rho(t) \quad . \tag{6}$$

It can be easily shown that $\Omega_{\mathrm{m}}(t) = 2q(t)$ in a Universe without cosmological constant, so that the two latter parameters are proportional and thus one only needs to determine one of the two.

Let us denote the current value of the density parameter as Ω_{m}. The cosmological parameters are tightly related to the geometry of the Universe. Indeed, it can be shown that:

$$kc^2 = H_0\, a_0[\Omega_{\mathrm{m}}(t) - 1] \quad , \tag{7}$$

where k is the spatial curvature and a_0 the present-time scale factor. It follows quite easily from this relation that:

- if $\Omega_{\mathrm{m}}(t) = 1$, then $k = 0$ and the Universe is described by the Einstein-De Sitter model;
- if $\Omega_{\mathrm{m}}(t) > 1$, then $k > 0$ and the Universe is described by the closed model;
- if $\Omega_{\mathrm{m}}(t) < 1$, then $k < 0$ and the Universe is described by the open model.

Hence, getting constraints on the cosmological parameters means constraining the cosmological model too.

6 THE ΛCDM MODEL

The standard adopted cosmological model is the so-called Λ-Cold Dark Matter (ΛCDM) model. The ΛCDM model is one of the most successful theories in Physics, since it has been supported by many observational evidences and has been able to explain all the cosmological observations performed so far.

Such model is named after:

- the cosmological constant Λ;
- the nature of dark matter in the Universe. A detailed treatment of this topic goes well-beyond the scope of this work. We will limit to mention that dark matter is nowadays regarded to be Cold Dark Matter, i.e. it is thought to be made of heavy, massive, slow particles, such as e.g. the WIMPs.[1]

The ΛCDM model assumes General Relativity as gravity theory at all scales, i.e. that the evolution of the Universe is gravitationally ruled by GR. Therefore, the growth of initial density perturbations and the further non-linear evolution, which lead to gravitationally-collapsed structures on different scales, must be accounted for by GR.

This cosmological model predicts that the Universe was birth in the so-called *Hot Big-Bang*, experienced an initial radiation-dominated phase until $z \sim 3000$, then a matter-dominated phase for $z \lesssim 0.5$ and that from $z \sim 0.5$ onward the Universe has entered a phase of accelerated expansion.

The concept of accelerated expansion of the Universe has been presented in Perlmutter et al. (1999); Riess et al. (1998). Such discovery was awarded with the *Nobel prize* to the authors in 2011. In such works the Universe is found to be accelerating its expansion since observed type Ia Supernovae appeared to be more distant than they should be in any of the cosmological models, even in a model with $\rho_m = 0$, i.e. a massless Universe (Milne's Universe). The most popular explanation which allows to find an agreement between theory and data is to introduce an energetic component, commonly called *Dark Energy*, which exerts a negative pressure, keeps its energy density constant across cosmic time and comes to dominate energetically in the late Universe, at $z \sim 0.5$.

A detailed treatment of this aspect goes beyond the scope of this work, but it is worth saying that the cosmological constant Λ has been commonly adopted to represent the accelerating factor in the expansion of the Universe. This effect can accounted for by adding one term to the Friedmann equations:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{a^2} + \frac{\Lambda}{3} \tag{8}$$

[1] Weakly Interactive Massive Particles

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda}{3} \quad , \tag{9}$$

where Λ has units [time]$^{-2}$ and represents an additional repulsive force. As mentioned above, these equations can be directly derived from the field equations of General Relativity and the addition of the Λ-term makes them covariant.

One can now define a proper density parameter for Λ as

$$\Omega_\Lambda(t) = \frac{\Lambda}{3\,H(t)} \quad . \tag{10}$$

Therefore the new overall density parameter can be expressed as

$$\Omega(t) = \Omega_\mathrm{m}(t) + \Omega_\Lambda(t) \quad . \tag{11}$$

Let us denote from now on the present-day density parameters as Ω_m and Ω_Λ. The geometry-parameters relation given in equation 7 can hence be re-written as

$$kc^2 = H_0\, a_0[\Omega_\mathrm{m}(t) + \Omega_\Lambda(t) - 1] \quad , \tag{12}$$

where the spatial flatness of Universe now requires $\Omega_\mathrm{m}t) + \Omega_\Lambda(t) = 1$.

Such definition can be generalized to account for additional energetic components in the Universe, such as e.g. radiation and neutrinos, as

$$\Omega = \sum_i \Omega_i + \Omega_\mathrm{k} \quad , \tag{13}$$

where in general Ω_i is the density parameter of the i-th component of the cosmological fluid and Ω_k is the density parameter of the spatial curvature. Data by Planck Collaboration et al. (2018) fixes $\Omega_\mathrm{k} = 0.0007 \pm 0.0019$, meaning that our Universe is today in a condition of high spatial flatness. The same reference fixes also $\Omega_\mathrm{m} = 0.315 \pm 0.007$ and $\Omega_\Lambda = 0.685 \pm 0.007$.

7 STRUCTURE FORMATION AND HIERARCHICAL CLUSTERING

As mentioned, the hypothesis of homogeneity of the Cosmological Principle is valid typically on scales $l > 100\,\mathrm{Mpc}$, but breaks down at smaller scales.

On (tens of) Mpc scales, matter forms an inhomogeneous pattern, called *cosmic web*, in which one finds filamentary overdense regions connected through knots, while other regions are left underdense. The cosmic web is actually observed at low redshifts and is regarded to be the result of the growth of the primordial perturbations in the initial density field across

cosmic time, of their non-linear evolution and ultimately of gravitational collapse ad clustering processes. The well-established *hierarchical clustering scenario* states that small structures form first and then cluster together assembling larger structures (see e.g. Bardeen et al., 1986; Fry & Peebles, 1978; Press & Schechter, 1974; White & Rees, 1978). The *Nature* paper Bond et al. (1996) showed that the preferential structures which form as a results of gravitational collapse are filaments and coined the expression cosmic web. In chapter 2, we will perform a more detailed and quantitative analysis of the cosmic web.

The methods and the estimators commonly adopted to measure the clustering of matter (galaxies, dark matter halos or just overdense regions of continuous fields) relies on measuring the statistical correlation of overdense regions on different scales and can be used either in configuration or in Fourier space.

The details are presented in what follows.

8 2-POINT CORRELATION FUNCTION AND POWER SPECTRUM

Let $\rho_r(\vec{x})$ be the matter density at coordinates $\vec{x}$ and smoothed on scale r , ρ_b be the background matter density, i.e. the average matter density in all the considered volume.

We define the overdensity field, evaluated at coordinates $\vec{x}$ and smoothed on scale r, as

$$\delta_r(\vec{x}) = \frac{\rho_r(\vec{x}) - \rho_b}{\rho_b} \quad . \tag{14}$$

One of the most common way to measure the clustering of the overdensity field is to look at the n-point statistical distribution in configuration space, defined by the so-called *n-point correlation function*. The n-point correlation functions (and their Fourier transforms, which will be introduced later in this section) constitute the so-called *summary statistics*. The study of summary statistics can be performed at various orders, but the most commonly-investigated orders are 2-point and 3-point statistics.

The 2-point correlation function (2PCF) is defined as

$$\xi(r) \equiv \langle \delta(\vec{x}), \delta(\vec{x} + \vec{r}) \rangle \quad , \tag{15}$$

where r is the magnitude of the separation vector $\vec{r}$.

The 2PCF basically estimates how likely is to find the same contrast density at points $\vec{x}$ and $\vec{x} + \vec{r}$. In other words, it describes the excess probability of finding two galaxies separated by this distance with respect to a uniform spatial distribution.

The 2PCF can be re-written in terms of Fourier transforms as

$$\xi(r) = \frac{1}{(2\pi)^6} \int \mathrm{d}^3k \int \mathrm{d}^3k' \, \langle \hat{\delta}(k), \hat{\delta}(k') \rangle \, e^{i\vec{k}\cdot(\vec{x}+\vec{r})} \, e^{i\vec{k}'\cdot\vec{x}} \quad . \tag{16}$$

We define the *power spectrum* $P(k)$ by linking it to the Fourier-space density perturbations through the relation

$$\langle \hat{\delta}(k), \hat{\delta}(k') \rangle \equiv (2\pi)^3 \, P(k) \, \delta_D^{(3)} \, (\vec{k} + \vec{k'}) \quad , \tag{17}$$

where $\delta_{\mathrm{D}}^{(3)}$ is the 3-dimensional Dirac delta. By substituting this relation in equation 16, making the Dirac delta act and computing the integral in the variable k' (getting 1 as a result), we finally find the relation

$$\xi(r) = \frac{1}{(2\pi)^3} \int \mathrm{d}^3k \, P(k) \, e^{i\vec{k}'\cdot\vec{x}} \quad . \tag{18}$$

This relation clearly shows that the power spectrum is the Fourier transform of the 2-point correlation function. This important result is also known as *Wiener-Khintchine* theorem. The power spectrum $P(k) \propto \langle |\hat{\delta}(k)|^2 \rangle$, i.e. at each wavenumber k it is proportional to the mean amplitude of plane waves of wavenumber k. This means that the power spectrum tells us how much a density fluctuation at scale k contributes to form the generic contrast density $\delta(\vec{x})$, i.e. tells us the power of the fluctuation at scale k. The power spectrum represents the measure of 2-point statistics in Fourier space.

9 BISPECTRUM

The 3-point statistics is measured in Fourier space by the bispectrum $B(k)$, which is defined as the ensemble average (see e.g. Scoccimarro et al., 1998; Sefusatti et al., 2006)

$$\langle \delta(\vec{k}_1), \delta(\vec{k}_2), \delta(\vec{k}_3) \rangle \equiv \delta_{\mathrm{D}}^{(3)}(\vec{k}_{123}) \, B(k_1, k_2, k_3) \quad , \tag{19}$$

where $\vec{k}_{123} = \vec{k}_1 + \vec{k}_2 + \vec{k}_3$ and δ_k represents the overdensity at scale k in Fourier space. The bispectrum is the Fourier-transform of the 3-point correlation function. Since we are now in Fourier space, let us recall that small k means large scales, large k means small scales. The presence of the Dirac delta ensures that we are accounting for closed triangular configurations, for which $\sum_i \vec{k}_i = 0$. Thus, the bispectrum gives an estimation of the number of Fourier-space closed triangles of sides k_1, k_2, k_3 or, alternatively, with two sides equal to k_1, k_2 and θ_{12} the angle subtended by the two vectors.

We then define the reduced bispectrum as

$$Q(k_1, k_2, k_3) = \frac{B(k_1, k_2, k_3)}{P(k_1)P(k_2) + P(k_1)P(k_3) + P(k_2)P(k_3)} \equiv Q(k_1, k_2, \theta_{12}) \quad . \tag{20}$$

The 3-point statistics will be studied throughout this work by means of the reduced bispectrum $Q(\theta_{12}, k_1, k_2)$.

10 SUMMARY STATISTICS ESTIMATORS

In order to compute correlation functions, power spectrum and bispectrum, proper estimators must be adopted.

10.1 2-point correlation function

One popular estimator for the 2PCF is the *Landy-Szalay* estimator (Landy & Szalay, 1993)

$$\xi_{\rm LS}(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)} \quad , \tag{21}$$

where, given a data sample with n_d objects and a spatially-uniform random sample with n_r objects, $DD(r)$, $RR(r)$ and $DR(r)$ are normalized sets of numbers of pairs as a function of the binned co-moving separation r, respectively within the data sample (DD), within the random sample (RR) and among the two samples (DR, cross-correlation).

10.2 Power spectrum

The estimator for the power spectrum adopted in this work is the so-called *Feldman-Kaiser-Peacock* estimator (FKP, Feldman et al., 1994).

Such estimator is based on three important assumptions (see Peebles, 1980):

- the distribution of the mass tracer we want to study can be regarded as the result of a Poisson sampling of an underlying continuous tracer distribution;
- at a given mode in Fourier space, the galaxy density field is characterized by fluctuations which are the result of a stochastic process characterized by a Gaussian distribution;
- the sample subtend such a small angle in the sky that the distant-observer approximation is valid.

Under these assumptions we can write the weighted galaxy (or equivalently halo, or any tracer) fluctuation in the region around point $\vec{r}$, according to the smoothing scale, as

$$F(\vec{r}) = \frac{\omega(\vec{r})}{\mathcal{N}^{1/2}}(n_g(\vec{r}) - \alpha n_s(\vec{r})) \quad , \tag{22}$$

where $n_g(\vec{r})$ corresponds to the observed number density of galaxies, $n_s(\vec{r})$ the same object for a synthetic catalog with randomly uniformly distributed particles (i.e. unclustered particles), $\omega(\vec{r})$ is a proper weighting function and α is a parameter which matches the number densities of the two catalogs, i.e. $n_s(\vec{r}) = \alpha^{-1} n_g(\vec{r})$. The number density of the real catalog is defined as $n_g(\vec{r}) = \sum_i \delta_{\rm D}(\vec{r} - \vec{r}_i)$, where $\delta_{\rm D}$ denotes a Dirac delta; the number density of the random catalog n_s is defined similarly. $\mathcal{N}$ is a proper normalization constant, defined as

$$\mathcal{N} = \int \mathrm{d}^3 r\, \bar{n}^2(\vec{r})\, w^2(\vec{r}) \tag{23}$$

Let us denote with $F(\vec{k})$ the Fourier transform of the weighted galaxy fluctuation field $F(\vec{r})$. By squaring it, taking the expectation value and working out the resulting expression, one gets

$$\left\langle |F(\vec{r})|^2 \right\rangle = \int \frac{\mathrm{d}^3 k'}{(2\pi)^3}\, P(\vec{k'}) \left| G(\vec{k} - \vec{k'}) \right|^2 + \frac{(1+\alpha)}{\mathcal{N}} \int \mathrm{d}^3 r\, \bar{n}(\vec{r})\, w^2(\vec{r}) \quad , \tag{24}$$

where

$$G(\vec{k}) = \frac{1}{\mathcal{N}} \int \mathrm{d}^3 r\, \bar{n}(\vec{r})\, w^2(\vec{r}) \exp(-i\, \vec{k}' \cdot \vec{r}) \quad . \tag{25}$$

Let D be a length scale which characterizes the depth of a survey (the side of the box, in our case). Then, if we restrict the study to scales $|\vec{k}| \gg 1/D$ (fair-sample hypothesis), then one can write

$$\left\langle |F(\vec{k})|^2 \right\rangle \simeq P(\vec{k}) + P_{\rm shot} \quad , \tag{26}$$

i.e. the raw power spectrum $|F(\vec{k})|^2$ is the sum of the true power spectrum and a shot-noise component. As a result, the true power spectrum can then be written as

$$\hat{P}(\vec{k}) = |F(\vec{k})|^2 - P_{\rm shot} \quad . \tag{27}$$

$P_{\rm shot}$ is the Poisson shot-noise component, defined as

$$P_{\rm shot} = \frac{(1+\alpha) \int \mathrm{d}^3 r\, \bar{n}(\vec{r})\, w^2(\vec{r})}{\int \mathrm{d}^3 r\, \bar{n}^2(\vec{r})\, w^2(\vec{r})} \quad . \tag{28}$$

Finally, to get the 1-dimensional power spectrum $\hat{P}(k)$ we average over spherical shells in Fourier space:

$$\hat{P}(k) = \frac{1}{V_k} \int_{V_k} \mathrm{d}^3k' \, \hat{P}(\vec{k}') \quad , \tag{29}$$

where $V_k = 4\pi k^2 \Delta k$ and Δk in the thickness of the shell.

10.3 Bispectrum

The general estimator for the bispectrum is given by e.g. Scoccimarro (2015); Scoccimarro et al. (1998)

$$\hat{B}(k_1, k_2, k_3) \equiv \frac{k_f^3}{V_\mathrm{B}(k_1, k_2, k_3)} \int k_1 \, \mathrm{d}^3q_1 \int k_2 \, \mathrm{d}^3q_2 \int k_3 \, \mathrm{d}^3q_3 \, \delta_D(\vec{q}_{123}) \, \delta_{q_1} \delta_{q_2} \delta_{q_3} \quad , \tag{30}$$

where the integrations are performed on shells of size Δk centered on k_i and where $V_\mathrm{B} \approx 8\pi k_1 k_2 k_3 \Delta k^3$ is a normalisation factor counting the number of fundamental triangles, defined by $\vec{q}_1, \vec{q}_2, \vec{q}_3$ on the Fourier grid, in a given triangle bin defined by k_1, k_2, k_3. Δk denotes the thickness of the spherical shell in Fourier space.

Under the assumption of Poisson shot-noise, the shot-noise correction presented in Matarrese et al. (1997) is applied:

$$B(k_1, k_2, k_3) = \hat{B}(k_1, k_2, k_3) - \frac{1}{(2\pi)^3 \bar{n}} [P(k_1) + P(k_2) + P(k_3)] - \frac{1}{(2\pi)^6 \bar{n}^2} \quad , \tag{31}$$

where $\bar{n}$ is the mean number density of the tracers inside the box and $P(k)$ is the power spectrum corrected for shot-noise.

In this work, we make use of the algorithm for the bispectrum estimator described in Scoccimarro et al. (1998). Eventually, the aliasing correction presented in Sefusatti et al. (2016) (interlacing technique) is applied.

11 APPLICATION TO THIS WORK

In order to quantitatively assess the accuracy of the mock catalogs of tracer of the dark matter field and of the adopted bias models we will investigate in this work, we calculate the 2- and 3-point statistics in Fourier space, i.e., the power spectrum and the bispectrum, respectively. In principle, both Fourier and configuration space analyses are equivalent. However, in practice they are not the same because of the truncated range in which each of them are computed. Configuration space and Fourier space summary statistics are indeed linked by Fourier transforms, which in the practical case

of calculations over a limited cosmological volume (the cubic simulation box) are characterized by integrations which cannot be performed on an infinite range. Moreover, the Fourier space representation is more sensitive to small scale clustering, since different scales mix across k-Fourier modes. Hence, a small scale deviation can be better detected in a power spectrum or bispectrum representation, rather than in the 2-point or 3-point correlation functions.

2 | THEORETICAL BACKGROUND

In this chapter we introduce the theoretical background of the topic which are investigated in this thesis.

First, we introduce the cosmic web and the way it has been defined and studied in the literature and propose a new way to characterize it. Then, we introduce the halo bias and provide all the operative concepts and definitions which will be adopted in the following chapters. Finally, we provide some basic concepts regarding the intergalactic medium and the Lyman-α and describe some common approximations used in such field.

1 THE COSMIC WEB

The cosmic web is the overall pattern, made of filamentary structures connected through knots and large underdense regions, that matter forms on Megaparsec scales as a result of the growth of primordial perturbations of the density field, gravitational collapse and clustering of smaller structures.

In general, the cosmic web is formed by dark matter, the intergalactic medium and gravitationally collapsed objects, such as galaxies, galaxy clusters and dark matter halos. Therefore, one can in general consider the various components of the cosmic web separately. We focus at this point on the dark matter cosmic web, i.e. the portion of the cosmic web defined by the dark matter density field. Later in this section we will also briefly mention the baryonic components of the cosmic web.

As mentioned in chapter 1, the expression *cosmic web* has been introduced in the pioneering *Nature* paper by Bond et al. (1996). In such work the authors compared numerically different mechanisms of structure formation, namely top-down vs bottom-up scenarios, finding a preference for the second. In this scenario, filaments form first and are the dominant overdense cosmic environments. Moreover, the authors proposed a first cosmic web classification based on the eigenvalues λ_i of the tidal field tensor $\mathcal{T}$ (namely, the hessian matrix of the gravitational potential; it will be formally defined later in this chapter). They classified the cosmic environments by defining the *ellipticity* $e \propto (\lambda_3 - \lambda_1)$ and the *prolateness* $p \propto (\lambda_1 + \lambda_3 - 2\lambda_2)$, where $\lambda_1 < \lambda_2 < \lambda_3$. In this way, in a filament $\lambda_1 \sim \lambda_2$ and thus $e \sim -p$, while in sheets $\lambda_2 \sim \lambda_3$ and $e \sim p$.

A more modern cosmic web classification, based on this idea, has been presented in Hahn et al. (2007). Let $\phi(\vec{x})$ be the gravitational potential at point $\vec{x}$ and let us define the gravitational tidal field tensor as

$\mathcal{T}_{ij}(\vec{x}) = \partial_i \partial_j \phi(\vec{x})$. The tidal field tensor is in principle defined at point $\vec{x}$ and characterizes a small patch around that point. In a practical mesh discretization of the Universe, in which one grids the considered cosmic volume (e.g. a cubic box) and subdivides it in many cells, the dependence on $\vec{x}$ is turned into a dependence on the cell in which the point $\vec{x}$ is contained.

Let then $\lambda_1 < \lambda_2 < \lambda_3$ be the eigenvalues of $\mathcal{T}_{ij}$, defined patch by patch. The authors define therefore a region of space to belong to a:

- void, if all the eigenvalues are negative;
- sheets, if two eigenvalues are negative and one is positive;
- filament, if two eigenvalues are positive and one is negative;
- knot, if all the eigenvalues are positive.

This classification, commonly known as T-web, is based on the kinematical argument that negative eigenvalues correspond to motions outwards in the corresponding directions, while positive eigenvalues means motions inwards towards an attractor point. Even though this classification is meaningful on a kinematical basis, it is quite phenomenological. For instance, it does not guarantee that what this classification calls filament is really a filament. Indeed defining quantitatively what a filament is turns out to be non-trivial.

One major problem this cosmic web model suffers is the arbitrariness of the choice of the threshold for the eigenvalues, which is commonly set to $\lambda_{\text{th}} = 0$. This choice is motivated by the aforementioned kinematical argument, but subsequent works (e.g. Forero-Romero et al., 2009) suggested to relax the definition of the threshold, in order to better match the visual appearance of the cosmic web with the cosmic web classification. Indeed, the authors of such work argue that what appears visually as e.g. a filament is misclassified if the threshold is set to $\lambda_{\text{th}} = 0$. Instead, they argue that highering the threshold to $\lambda_{th} \sim 0.2$ allows to make the resulting web classification reproduce the visual impression of the cosmic web. Other works (Aragón-Calvo et al., 2007; Cautun et al., 2012) proposed a multi-scale cosmic web classification.

The perspective we have adopted in Kitaura et al. (2020) and in this work is that of relying on the invariants of the gravitational tidal field tensor, defined as:

- $I_1 = \lambda_1 + \lambda_2 + \lambda_3 \equiv \delta_{\text{L}}$, corresponding to the linearized density field;
- $I_2 = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3$, corresponding to the sum of the determinants of the minors;
- $I_3 = \lambda_1\lambda_2\lambda_3$, corresponding to the determinant;
- $I_4 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$;

- $I_5 = \lambda_1^3 + \lambda_2^3 + \lambda_3^3$.

We dub I-web the following ensemble of invariants: $\{I_1, I_2, I_3\}$. Notice that we can consider the full non-linear density field δ to be the sum of δ_L + higher order terms. Therefore, having the evolved non-linear density field available, we will likely adopt as I-web the set $\{\delta, I_2, I_3\}$ in the following chapters.

By performing some analytical calculations, I-web is found to be tightly connected to T-web. Let us assume, without loss of generality, that $\lambda_{th} = 0$. Then, one can find the following relations (see Kitaura et al., 2020, for the details of the derivation):

- a knot is characterized by $I_3 > 0, I_2 > 0, I_1 > \lambda_1$;
- a filament is characterized by $I_3 < 0, I_2 < 0$ or $I_3 < 0, I_2 > 0, \lambda_3 < I_1 < \lambda_3 - \lambda_2\lambda_3/\lambda_1$;
- a sheet is characterized by $I_3 > 0, I_2 < 0$ or $I_3 > 0, I_2 > 0, \lambda_1 - \lambda_2\lambda_3/\lambda_1 < I_1 < \lambda_1$;
- a void is characterized by $I_3 < 0, I_2 > 0, I_1 < \lambda_1$.

The choice of a threshold $\lambda_{th} \neq 0$ just shifts these relations. It may seem that one has to specify some of the eigenvalues in such relations. However, it can shown that, if the linearized density field I_1 is known, it is enough to know the signs of I_2 and I_3 to constrain the different cases and distinguish between them.

What has been exposed so far makes it clear that I-web consists in a re-definition of the cosmic web, connected to the T-web definition commonly adopted in the literature. However, the I-web is found to improve the T-web in two important aspects:

- first, I-web does not suffer the arbitrariness of the choice of the threshold, which is crucial to distinguish between different cosmic environments in the T-web;
- second, T-web describes the information regarding the cosmic environment in which a certain region is found in just 4 categories (knots, filaments, sheets, voids). Instead, I-web opens and exploits such information in an arbitrarily high number of bins, provided that the invariants of the tidal field tensor span continuously the parameter space.

These two important features allow I-web to get rid of the phenomenological choice of the threshold and lead to a net gain of information with respect to T-web. This aspect will turn out to be crucial in chapter 4, in the scope of the study of halo bias.

Finally, it is worth mentioning two final points.

First, in the context of galaxy formation and evolution, a number of correlations between galaxy properties and cosmic environments (assume T-web) in which they reside are found. For instance, galaxy spins are found to be aligned with the filaments in which they live. Or again, spiral galaxies are likely not to be found in knots, while elliptical galaxies are preferentially found there. We believe these kind of studies could take advantage of using the I-web perspective instead of the T-web one. The former would indeed allow to expand the 4 cosmic environments of the latter into the multidimensional continuous distribution of the invariants of the tidal tensor. Hence, the regions of the parameter space spanned by $\{I_1, I_2, I_3\}$ and the considered galaxy properties could be a useful site where to look at to perform these studies.

Second, as mentioned previously the components of the cosmic web can in principle be investigated separately. In other words, one can think of performing a cosmic web classification based not on the dark matter density field but on one of its tracer fields, when this makes sense. This is important because some baryonic components of the cosmic web, such as the ionized gas contained in the IGM, might be sensitive to small-scale non-linear processes which involve baryons but not dark matter. For instance, photo-heating and photo-ionization of the gas by photons of the UV background and radiative cooling directly involve baryonic particles and photons and regulate the relative abundance of neutral and ionized species, but do not affect significantly the spatial distribution of dark matter. For this reason, it may turn out to be interesting to consider also the characteristics of the cosmic web of some tracer fields, when possible. In chapter 5 we will show that the intergalactic ionized gas density has a very high correlation with the dark matter field at all scales, so that we can safely compute I-web from such field. Furthermore, we will show that this aspect is very important because the gas I-web will be found to constrain the spatial distribution of HI more accurately than the DM I-web, at the scales we will be investigating.

2 HALO BIAS

Dark matter halos consist in gravitationally collapsed objects made of dark matter, which form as a result of gravitational collapse, in a complex non-linear dependence on the dark matter density field. Pioneering works such as Press & Schechter (1974) started by considering a model of spherical collapse and computed an analytical expression for the mass function of such objects. Later on, Sheth et al. (2001) implemented a model of ellipsoidal collapse, which updates the spherical collapse model by assuming that a proto-halo collapses before in one direction and then in the others, leading to an ellipsoidal shape.

Of course, the spatial distribution and clustering of halos is tightly related to the underlying diffuse dark matter. In general, the gravitational collapse is favoured in the denser region of the dark matter field. However, due to the great complexity of the process which leads to the formation of halos, establishing a well-defined relation between the spatial distribution of halos and dark matter is non-trivial.

The *halo bias* describes, in a statistical sense, the relation between the distribution of halos and that of dark matter. This topic has been studied for decades and is very wide, comprising several theoretical and observational aspects (see Desjacques et al., 2018, for a recent review on halo/galaxy bias).

In this work we focus on the so-called *Eulerian bias*, which expresses the halo overdensity at a certain fixed cosmic time as a general functional dependence on the underlying dark matter distribution in Eulerian coordinates:

$$\delta_{\rm h}(\vec{r}) = F\left[\delta(\vec{r}), \mathrm{X}, \mathrm{Y}, \dots\right] \quad , \tag{32}$$

where $\delta_{\rm h}(\vec{r})$ and $\delta(\vec{r})$ are respectively the halo and dark matter overdensity at point of coordinates $\vec{r}$ and X, Y, ... denote possible more complex non-linear and non-local dependencies (i.e., which depend not only on the properties of the patch around $\vec{r}$, but also on the properties of the surrounding regions) on the dark matter field or on its velocity field.

The halo bias has been classically investigated adopting a Perturbation Theory (PT hereafter) approach. Unlike the CMB, which is accurately described by linear (first-order) PT, the perturbations in the matter density of the Universe at redshifts $z \lesssim 10$ are non-linear. However, the degree of non-linearity depends on the scales considered. One can roughly divide Large-Scale Structure into two regimes: large scales, which can be regarded as quasi-linear, where PT converges to the correct result if carried out to sufficiently high order; small scales, which are highly non-linear and cannot be described by PT.

Let us focus for a moment on large scales, where PT applies. In this regime, the number (density) of halos in a certain patch of the Universe is expressed as a Taylor-expansion of the (local) dark matter overdensity field:

$$\delta_{\rm h}(\vec{r}) = F\left[\delta(\vec{r})\right] = \sum_n \frac{b_n}{n!}\left[\delta(\vec{r})\right]^n \quad , \tag{33}$$

where b_n denote the Taylor coefficients of the expansion. Such example of expansion accounts only for local terms, i.e. first and higher order powers of the dark matter overdensity δ, but in a more general framework might account also for non-local dependencies which describe the degree of anisotropy of the environment in which the point $\vec{x}$ is located, the cosmic web type in which it resides and the underlying geometry of the gravitational potential. One important reference paper for halo bias study with the PT approach is McDonald & Roy (2009), in which the halo bias terms

are constructed from the tidal field tensor $\mathcal{T}$ and the velocity shear. In, other significant works (e.g. Crocce & Scoccimarro, 2006; Werner & Porciani, 2019), the author investigate how to properly renormalize cosmological PT.

The assessment of the bias of dark matter halos has also been performed through the large-scale signal of two- points statistics, either the correlation function or the power spectrum (see e.g. Tinker et al., 2010). However, some works have explored the possibility of using the bispectrum to perform a measure of the halo bias on large scales (e.g. Matarrese et al., 1997). These approaches formalize a large-scale bias which can be characterized as a function of intrinsic properties such as the halo mass, concentration or spin. From a phenomenological point of view, the halo bias has a dependency on properties of the dark matter field such as its tidal field, its formation time and a number of other dependencies which constitute the so-called *assembly bias* (Angulo et al., 2008; Contreras et al., 2019; Gao & White, 2007; Obuljen et al., 2019; Zentner, 2007, and references therein).

The bias is nowadays regarded to have a stochastic scale-dependent nature. Indeed, while large scales can be regarded as quasi-linear, on intermediate and small scales highly non-linear processes become important and regulate the collapse and the formation of structures.

Throughout this work, the operative definition of bias implemented in our analysis is a statistical stochastic bias, inspired by Dekel & Lahav (1999). In this picture, the functional dependence described in equation 32 is thought as a conditional multi-variate probability distribution, which expresses the probability of finding a certain halo overdensity at coordinates $\vec{r}$, given the dark matter density evaluated at $\vec{r}$ and all the other characterizing dependencies:

$$\delta_{\rm h}(\vec{r}) \leftharpoondown P\left(\delta_{\rm h}(\vec{r})\right| \delta(\vec{r}), \mathrm{X}, \mathrm{Y}, \dots) \ . \tag{34}$$

The key idea applied in chapter 4 is including in the stochastic bias the following dependencies:

$$\delta_h(\vec{r}) = F(\delta, \mathcal{T}, \Gamma, \Sigma, \omega, \epsilon) \quad , \tag{35}$$

where, beside local dependence on δ, $\mathcal{T}$ denotes the long-range non-local dependence on the gravitational tidal field tensor tidal field, Γ includes short-range non-local bias terms, ω and Σ represent respectively the vorticity (curl) and the shear of the velocity field and ϵ a white noise terms in the power spectrum measurement (Feldman et al., 1994). This last noise term accounts for the fact that the halo field is a discrete realization of the expected number counts of objects per volume element. This is responsible for a stochastic uncertainty and can hence be modelled with a shot-noise term. Actually, we will restrict the functional dependence to local δ and long-range non-local terms $\mathcal{T}$, leaving terms built from the velocity field and short-range non-local terms for future works.

The bias model presented in equation 35 is inspired by the bias expansion presented in McDonald & Roy (2009) (McD-R, hereafter), which formalizes a PT expansion up to 3rd order, including the dependence on the tidal field tensor.

Notice that the purpose of this work is not treating the bias in PT. The aim is to formalize and test a stochastic bias model which includes the dependence on terms at different orders, to establish which dependencies are crucial to describe the bias on some scales and which are instead negligible. Therefore, one has to have clear which are the terms which constitute the different perturbative orders. To this end, we try to update the expansion of McD-R, accounting explicitly for all the terms up to 3rd order, including short-range non-local terms (already discussed in McD-R, but we generalize them here) and velocity terms.

Let us re-define the traceless tidal field tensor as

$$s_{ij}(\vec{r}) \equiv \mathcal{T}_{ij} - \frac{1}{3}\,\delta^{\mathrm{K}}_{ij}\,\delta_{\mathrm{L}}(\vec{r}) \quad , \tag{36}$$

where δ_{L} is the linearized density field and δ^{K}_{ij} is the Kronecker delta. Let us then build the following scalar terms[1] based on s_{ij} as

$$s^2 \equiv s_{ij}\,s_{ij} \quad , s^3 \equiv s_{ij}\,s_{jk}\,s_{ki} \quad , \tag{37}$$

which are non-local terms at respectively 2nd and 3rd order.

McD-R and the majority of the works in the literature perform perturbative expansions on very large scales and consider the linearized density field. Here, instead, we are considering the full non-linear density field defined on few Mpc scales. According to the Helmoltz decomposition, one can express the velocity field as the sum of a divergence-free and a curl-free vector fields. Going down to small scales, the divergence-free component of the velocity becomes increasingly important and relevant on sub-Mpc scales. However, according to Kitaura et al. (2012), on scales $l \sim 5\,\mathrm{Mpc}$, at which we are smoothing the density field, the divergence-free component is approximately negligible with respect to the curl-free component. Therefore, we can consider local and non-local terms built from the density field (i.e., all the first line in equation 38) as equivalent to terms built from a curl-free velocity field with $\theta = \delta$, where $\theta = \nabla \cdot \vec{v}$ is the weighted velocity divergence and $\theta = \delta$ in linear perturbation theory.

Then, we notice that we can build shear terms $\Sigma_{ij} \sim \partial_i v_j$ from both the curl-free and the divergence-free part of the velocity field, dubbed respectively $F_{\mathrm{shear}}(\vec{v}(\vec{r})|_{\,\mathrm{curl\text{-}free}})$ and $F_{\mathrm{shear}}(\vec{v}(\vec{r})|_{\,\mathrm{div\text{-}free}})$. The curl-free part has vanishing curl (the curl of a curl-free field is vanishing), but the divergence-

[1] Because we are assuming homogeneity and isotropy, the bias parameters can be only constant scalars.

free component has non-zero curl, from which we can build additional terms $F_{\rm curl}(\vec{v}(\vec{r}))$.

Finally, we group in $F_{\delta^{\rm SR}}(\partial_i^l\,\partial_j^l\,\delta(\vec{r}))$ the short-range terms and in $F_\epsilon(\epsilon(\vec{r}))$ the noise terms.

The resulting Taylor-expansion in Eulerian coordinates hence is:

$$\delta_{\rm h}(\vec{r}) = \underbrace{\overbrace{c_\delta\delta(\vec{r})}^{\text{local}}}_{\text{first order}} + \underbrace{\overbrace{\frac{1}{2}c_{\delta^2}\left(\delta^2(\vec{r})-\langle\delta^2\rangle\right)}^{\text{local}} + \overbrace{\frac{1}{2}c_{s^2}\left(s^2(\vec{r})-\frac{2}{3}\langle\delta^2\rangle\right)}^{\text{non-local}}}_{\text{second order}} + \underbrace{\overbrace{\frac{1}{3!}c_{\delta^3}\delta^3(\vec{r})}^{\text{local}} + \overbrace{\frac{1}{2}c_{\delta s^2}\delta(\vec{r})s^2(\vec{r}) + \frac{1}{3!}c_{s^3}s^3(\vec{r})}^{\text{non-local}}}_{\text{third order}} + \underbrace{\overbrace{\mathcal{O}(F_{\delta,\mathcal{T}}(\delta)|^4)}^{\text{local \& non-local}}}_{\text{fourth order+}}}_{\text{curl-free \& }\theta=\delta\text{ terms}}$$

$$+ \overbrace{\underbrace{F_{\delta^{\rm SR}}(\partial_i^l\partial_j^l\delta(\vec{r}))}_{\substack{\text{short range }\theta=\delta\text{ terms }l\in\mathbb{N}\\ \text{first order+}}} + \underbrace{F_{\rm shear}(\vec{v}(\vec{r})|_{\rm curl-free})}_{\theta\neq\delta\text{ terms}} + \underbrace{F_{\rm shear}(\vec{v}(\vec{r})|_{\rm div-free}) + F_{\rm curl}(\vec{v}(\vec{r}))}_{\text{vorticity }\theta=\delta\ \&\ \theta\neq\delta\text{ terms}}}^{\text{non-local}} + \underbrace{\overbrace{F_\epsilon(\epsilon(\vec{r}))}^{\text{local \& non-local}}}_{\substack{\text{noise terms}\\ \text{first order+}}} \quad ,$$
$$\text{third order+}$$

where the cs are some bias factors and the braces specify the locality or non-locality of the terms and their perturbative order.

With this formula in mind, in chapter 4 we will build a number of numerical bias models including terms of different orders and test them in the perspective of assessing which are the crucial dependencies on large scales.

3 IGM AND LYMAN-α FOREST

The intergalactic medium (IGM hereafter) is a fundamental component of our Universe. It is made of all the baryonic matter which does not belong to collapsed structures, i.e. matter which does not belong to galaxies or galaxy clusters.

The IGM represents one of the fundamental component of the cosmic web, together with galaxies, galaxy clusters, halos and diffuse dark matter. Its main constituents are hence expected to follow a spatial distribution which is tightly related to the dark matter cosmic web, in such a way that we can regard them as biased tracers of the dark matter field.

In general, the formation and evolution of galaxies are intimately connected to the IGM, because the diffuse gas of the IGM represents the main fuel replenishing gas reservoirs during gas accretion processes in galaxies (see Popping et al., 2015, and references therein).

In order to understand what the IGM is made of, let us neglect in first approximation all the metals and consider the Universe as made exclusively of H and He. At redshift $z \lesssim 6$, H is regarded to have completed the Reionization process and therefore it is present mostly in the form of ionized hydrogen, leaving however a non-negligible residual of neutral H. Equivalently, the same can be said for He at redshift $z \lesssim 3$. Because in

principle H is by far more abundant than He, let us neglect neutral atomic He and group, for simplicity, the gaseous constituents of the IGM into neutral H (HI hereafter) and ionized gas, the latter taking into account all the ionized gas species.

A common assumption adopted in the literature is to regard that the IGM is in photo-ionization equilibrium. This approximation predicts that photo-heating and photo-ionization processes due to absorption of photons of the UV background are challenged and balanced by radiative cooling. This condition is mathematically expressed by the following relation, dubbed *photo-ionization equilibrium equation*

$$n_{\rm HI} \int_{\nu_{\rm L}}^{\infty} \frac{4\, J_\nu\, \sigma_{\rm H}(\nu)}{h\nu}\, {\rm d}\nu = n_e\, n_P\, \sigma_{\rm A}(T) \quad , \tag{38}$$

where $n_{\rm HI}$ is the HI number density, $\nu_{\rm L} = 13.6$ eV is the energy threshold for photo-ionization, $\sigma_{\rm H}(\nu)$ is the photo-ionization cross-section, expressed as $\sigma_{\rm HI}(\nu) = \sigma_{\rm L}\,(\nu/\nu_{\rm L})^{-3}$ and $\sigma_{\rm L} = 6.3 \times 10^{-18}{\rm cm}^2$, J_ν is the mean intensity of the ionizing radiation (in units of energy per unit area, time, solid angle, and frequency interval, i.e. is a flux), h is the Planck constant, n_e and n_p are the number densities of electrons and protons and $\sigma_{\rm A} = \sum \langle \sigma_n\, v_e \rangle$ is the radiative recombination coefficient, i.e. the product of the electron capture cross-section σ_n and the electron velocity v_e, averaged over a thermal distribution and summed over all atomic levels n. Notice that $\sigma_{\rm A}$ is a function of the temperature. The left-hand side of the equation expresses the photo-ionization rate (related to photo-heating) per unit volume, while the right-hand side of the equation stands for the rate of radiative recombinations (related to radiative cooling). Therefore, the equation expresses a balance between photo-ionization and radiative recombination rates and holds in an optically thin, pure HI medium with number density $n_{\rm HI}$. The photo-ionization equation can be written also for He, with a proper frequency threshold for ionization ($E > 24.6\,$eV for first ionization, $E > 54.4\,$eV for second ionization).

Another common approximation used to treat the IGM consists in assuming a tight power-law relation between the temperature T and the gas density ρ:

$$T = T_0 \left(\frac{\rho}{\bar{\rho}}\right)^{\alpha} \quad , \tag{39}$$

where $\bar{\rho}$ is the mean gas density. T_0 is an IGM temperature scale which depends on the Reionization history of the Universe and on the spectral shape of the UV photon background and lies typically in the interval $4000\,{\rm K} \lesssim T_0 \lesssim 10000\,{\rm K}$. According to Hui & Gnedin (1997), $0.3 \lesssim \alpha \lesssim 0.6$.

These two approximations, combined together, allow to express the optical depth τ for the Lyman-α absorption (see below in this section) as a function of the ionized gas overdensity. Such relation, called *Fluctuating*

Gunn-Peterson approximation (FGPA, see e.g. Cieplak & Slosar, 2016; Gunn & Peterson, 1965; Seljak, 2012; Weinberg et al., 1997) can be written as

$$\tau = A(z)(1+\delta)^{\alpha} \quad , \tag{40}$$

where $\alpha = 2 - 0.7(\gamma - 1)$, $\gamma - 1 = \mathrm{d}\log\rho/\mathrm{d}\log T$ (a typical value is $\alpha \sim 1.6$, with $\alpha = 2$ being the isothermal case) and δ is the gas overdensity. $A(z)$ is a constant, given by the expression (Weinberg et al., 1997):

$$A(z) = 0.835 \left(\frac{1+z}{4}\right)^{6} \left(\frac{\Omega_b h^2}{0.02}\right)^{2} \left(\frac{h}{0.65}\right)^{-1} \times \\ \times \left(\frac{H(z)/H_0}{4.46}\right)^{-1} \left(\frac{\Gamma_{\mathrm{HI}}}{10^{-12}s^{-1}}\right)^{-1} \left(\frac{T_0}{10^4\mathrm{K}}\right)^{-0.7} \quad ,$$

where z is the redshift, Ω_b is the baryon density parameter, $H(z)$ and H_0 are respectively the Hubble parameter and the Hubble constant, Γ_{HI} is the HI photo-ionization rate.

A great interest and a wealth of studies, both theoretical and observational, have been dedicated to HI. This is largely justified by the fact that HI is a very efficient tracer of the matter distribution, both in absorption and in emission. Indeed, HI is responsible for both the so-called Lyman-α forest absorption and for the 21-*cm emission line*. Even though a detailed treatment of the radio-emission of neutral hydrogen goes well beyond the scope of this work, it is worth mentioning that current and future observational campaigns carried out with radio-telescope arrays (one among all, the Square Kilometre Array) will perform a detailed mapping of HI in galaxies and in the IGM up to high redshift, opening a new important observational window in the Universe.

The Lyman-α forest consists in a thick ensemble of absorption lines which characterize spectra of distant quasars. Such lines are originated by the Lyman-α transition of HI ($\lambda = 1216$Å), which occurs when energetic photons ($\lambda < 1216$Å) encounter intergalactic clouds containing HI along the line-of-sight. Due to cosmological redshift, photons which are absorbed by HI at different distances from the observer produce lines at different redshifts in the final observed spectra. This is why, even though the absorption process occurs always at the same wavelenght, the final signature in quasar spectra is an ensemble of discrete spaced absorption lines.

The Lyman-α forest plays a major role in Cosmology because it traces efficiently the matter distribution in a redshift domain which is not well investigated by galaxy surveys, i.e. $z \gtrsim 2$. Figure 2, extracted from Tom Theuns's talk at Durham University in 2008, shows a clear example of the observed Lyman-α forest. The upper and lower panels show respectively a spectrum of a nearby quasar and of a distant quasar. One clearly notices that the spectrum of the distant quasar is characterized by a set of thick

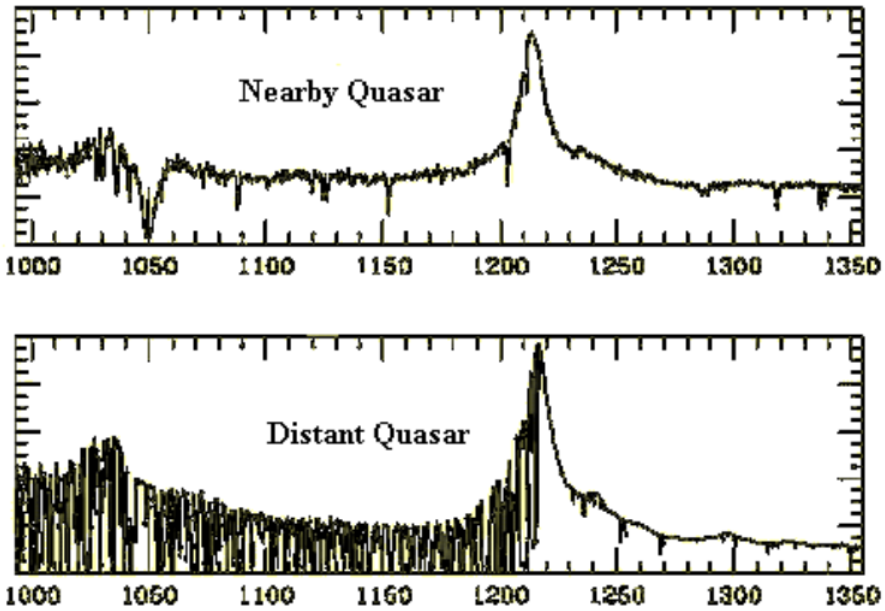

Figure 2: A comparison between real observed spectra of a nearby quasar (upper panel) and a distant quasar (lower panel). One clearly notices that the spectrum of the distant quasar presents a thick set of absorption lines, the so-called Lyman-α forest, in the wavelength region $1000\,\text{Å} < \lambda < 1200\,\text{Å}$. Figure extracted from Tom Theuns's talk at Durham University in 2008.

absorption lines (the Lyman-α forest) in the wavelength region $1000\,\text{Å} < \lambda < 1200\,\text{Å}$.

Even though in principle the Lyman-α forest consists in a discrete set of absorption lines, Weinberg et al. (1997) suggested to regard the Lyman-α forest as a continuous non-linear map of the line-of-sight density field. In fact, if one imagined to observe an infinite number of light rays coming from quasars, along an infinite number of parallel line-of-sights, than the Lyman-α forest could be really thought as a continuous field in a volume, as much as matter density. What one does in practice is calculating the optical depth τ for the Lyman-α transition (see chapter 5 for the details about the computation of τ) and then obtain the transmitted Lyman-α fluxes F by taking $F = \exp(-\tau)$. In the end, such transmitted Lyman-α forest fluxes can be considered as continuous biased tracers of the underlying matter density field.

On large scales, the Lyman-α forest bias can be treated in linear theory and one can link its 3-dimensional power spectrum to the matter power spectrum by means of the relation

$$P_F(\vec{k}) = b_\delta^2(1 + \beta\mu^2)^2 P_\delta(\vec{k}) + P_N \quad , \tag{41}$$

where $P_\delta(\vec{k})$ is the matter power spectrum, $\beta = f\, b_\eta / b_\delta$, f is the logarithmic growth rate, b_δ and b_η are proper bias parameters and μ is the cosine of the

angle between the line-of-sight and the vector $\vec{k}$. This formula is equivalent to the Kaiser redshift-space distorsions (RSD, Kaiser, 1987) for galaxies, where however here β is determined not only by b_δ. P_N is a white noise term which is thought to be small and is often ignored. As proposed by McDonald & Eisenstein (2007) and later implemented in the BOSS survey, the Lyman-α power spectrum can be used as a powerful tool to measure the large-scale matter power spectrum and e.g. to measure the Baryonic Acoustic Oscillations (BAO) scale.

On the other hand, on small scales the Lyman-α power spectrum is affected by a number of highly non-linear complex processes which govern the evolution of the IGM, mainly related to star and galaxy formation and evolution. One common assumption employed to tackle this issue is to assume that the IGM is in a condition of photo-ionization equilibrium, presented previously in this section, which allows a simple analytical treatment of the Physics on small scales.

The part of this work (and future works) related to Lyman-α forest focuses on learning from a hydrodynamical simulation the bias relation (in a stochastic sense) with the underlying dark matter density field. This is done in order to extract information on the nature of such bias relation and to ultimately sample fast and accurate Lyman-α forest mock catalogs, both in real and in redshift space, in the view of DESI.

3 THE BIAS ASSIGNMENT METHOD

In this chapter we introduce the topic of mock catalogs production and present the BAM method, on which this work relies. As we have anticipated in chapter 2, our goal is to extract the optimal bias model between halos and the dark matter field and, more in general, to extract the bias relation between a field and any of its tracers. We will see that BAM is a very efficient machinery in this sense.

The *Bias Assignment Method* (BAM hereafter, Balaguera-Antolínez et al., 2018, 2019; Pellejero-Ibañez et al., 2020) is a method designed for creating fast and precise mock catalogs of dark matter tracers. It consists in a bias-mapping technique, based on learning the bias relation in a stochastic sense (see chapter 2) from a reference simulation and use such relation to perform a massive sampling of mock catalogs. BAM has been conceived to produce halo mock catalogs, i.e. to perform random samplings of a discrete field. In other words, in its original version it had been designed to assign number counts, i.e. integer numbers. However, such method has been extended to cope with continuous density fields and is now able to sample mock catalogs of a much wider range of physical quantities, such as the main properties of the intergalactic medium. This allows to realize that this method is very versatile and applicable to a variety of cosmological fields. Anyway, in order to acquire a sufficient understanding of how this method works and which are the theoretical assumptions it is based on, let us go through the details.

The construction of mock catalogs has become the standard approach to estimate errors on cosmological observables. This is due to the fact that we lack a reliable and accurate analytical model, accounting for the highly non-linear evolution of the density field, galaxy bias, baryon effects (Eisenstein & Hu, 1998), redshift-space distorsions (Kaiser, 1987) and systematic effects (e.g. survey geometry).

To this end, the first attempt to create mock catalogs relied on N-body simulations. However, due to the large computational time and memory storage usually required by current N-body simulations, researchers have started to seek another way to build mock catalogs since late 1990s. In order to speed up the generation of the mocks, a number of different methods has been proposed up to now. The assessment of the accuracy of a mock-making method is usually performed by comparing the results of the method, in terms of summary statistics of the created mock, to a reference N-body simulation. In particular, the power spectrum and eventually higher order summary statistics of the reference field extracted from the N-body

simulation are compared to the ones of the mock catalog, the latter obtained by evolving the same initial conditions of the simulation.

The most common application of such methods is the production of dark matter halo and galaxy mock catalogs. The first works aimed at producing galaxy/halo mock catalogs relied on predictive analytical methods employing approximate gravity solvers based e.g. on Lagrangian perturbation theory. Examples of these methods are e.g. the PINOCCHIO (Monaco et al., 2002) and PEAK-PATCH (Bond & Myers, 1996) algorithms. However, these methods turn out to be rather inaccurate on small scales, since the mock power spectrum suffers deviations of more than 5% with respect to the true reference power spectrum at such scales. Recently there have been improvements in fast gravity solvers, like e.g. ICE-COLA (Tassev et al., 2013) and FastPM (Feng et al., 2016). These gravity solvers are less severely affected by the aforementioned inaccuracies, but require very large memory facilities, comparable to those of N-body simulations. Furthermore, they turn out to be only slightly faster than the full N-body approach, so they still do not meet the requirements which are needed for a fast and massive production of mock catalogs.

Another class of methods is instead based on the mapping of the halo distribution by relying on an approximated dark matter density field and an analytical bias prescription, i.e. an analytical relation which expresses the halo overdensity as a functional of the dark matter density (see chapter 2). In practice, once one has the approximated density field and the halo bias relation, the halo mocks are sampled according with these two. Works which adopt this approach are e.g. PThalos (Scoccimarro & Sheth, 2002), MoLUSC (Sousbie et al., 2008), PATCHY (Kitaura et al., 2014), QPM (White et al., 2014), EZ-MOCKS (Chuang et al., 2015) and HALOGEN (Avila et al., 2014).

BAM relies basically on the same stochastic bias mapping idea employed by PATCHY, but does not assume any analytical prescription for the halo bias relation. Instead, the method learns the bias relation between the halo and dark matter density fields from a reference N-body cosmological simulation, calibrates the bias in a proper way and then uses such calibrated relation to do the sampling. Hence, differently from other methods, BAM is designed to be parameter-free. This aspect is very important, because it allows to capture all the non-linearities involved, without the need of any assumptions on free parameters or analytical formulas.

For simplicity we discuss first how BAM is applied to sample halo mock catalogs. Later on, in section 3, we will present how the method is generalized to create mock catalogs of continuous physical quantities.

Concretely, the bias relation is measured from an accurate dark matter density field (DMDF hereafter), result at a given redshift of the exact N-body gravity calculation, and the dark matter halo field (DMHF hereafter) obtained by applying a Friends-of-Friends halo-finder algorithm to the dark matter field. The bias is interpreted in a stochastic sense, in the spirit of

Dekel & Lahav (1999), as the probability of finding a certain number (density) of halos conditional to a set of properties of the DMDF. In other words, the bias is expressed as a multi-variate conditional probability distribution.

Once the bias relation is known, the code calibrates it with an iterative procedure and then uses it to map the spatial distribution of halos of different realizations. As will be explained in details later, this is possible thanks to approximated DMDFs resulting from the evolution of some Initial Conditions (ICs hereafter) with a gravity solver. In our case, we employ Augmented Lagrangian Perturbation Theory (ALPT), already adopted successfully in PATCHY (Kitaura et al., 2014).

BAM has been shown to be a very efficient method and is found to improve the other methods proposed in the literature. Indeed, it reproduces the 2-point statistics of the reference simulation typically with percent accuracy. Let R be the percentage residuals of the mock power spectrum with respect to the reference power spectrum, averaged over all the spherical shells in Fourier space, defined as

$$R = \left(\sum_i \frac{|P_{\rm mock}(k_i) - P_{\rm ref}(k_i)|}{P_{\rm ref}(k_i)} \right) / N_{\rm shells} \quad , \tag{42}$$

where the sum is performed over the spherical shells k_i in Fourier space and $N_{\rm shells} = N_{\rm cells}/2$ is the number of shells over which the sum is performed. Balaguera-Antolínez et al. (2018) and Balaguera-Antolínez et al. (2019) clearly show that BAM is able to sample halo mock catalogs whose $P(k)$ are characterized by $R \lesssim 1\%$ with respect to the reference $P(k)$. An example of such finding is shown in Figure 3.

Moreover, BAM is computationally very fast and requires modest computing facilities. E.g., the code is able to produce halo mock catalogs on a $N = 500^3$ cells mesh over a short time scale, typically $\Delta t \sim 100\,$s, and with a memory requirement $\sim 1 - 5\,$Gb. Such features are not commonly encountered in other methods (see e.g. Blot et al., 2019; Colavincenzo et al., 2019; Lippich et al., 2019, for a comparison of halo mock building methods). For this reason, BAM has all the right credentials to become a popular technique to produce large amounts of mock catalogs.

BAM adopts a mesh representation of the fields, i.e. it interpolates the fields on a grid with a certain number of cells. Therefore, when the output of a simulation is given in the form of a set of particles, one needs to interpolate such particles on a mesh with a mass assignment scheme (commonly Nearest-Grid Point (NGP) or Cloud-in-Cell (CIC), Hockney & Eastwood, 1981). Once the interpolation has been carried out, all the calculations in the code are performed on the mesh, cell by cell. Hence, one must keep in mind that all the computations presented in this work are performed with this approach and reason in terms of dealing with a grid of cells. More details about the representation of a set of particles on a mesh, the mass

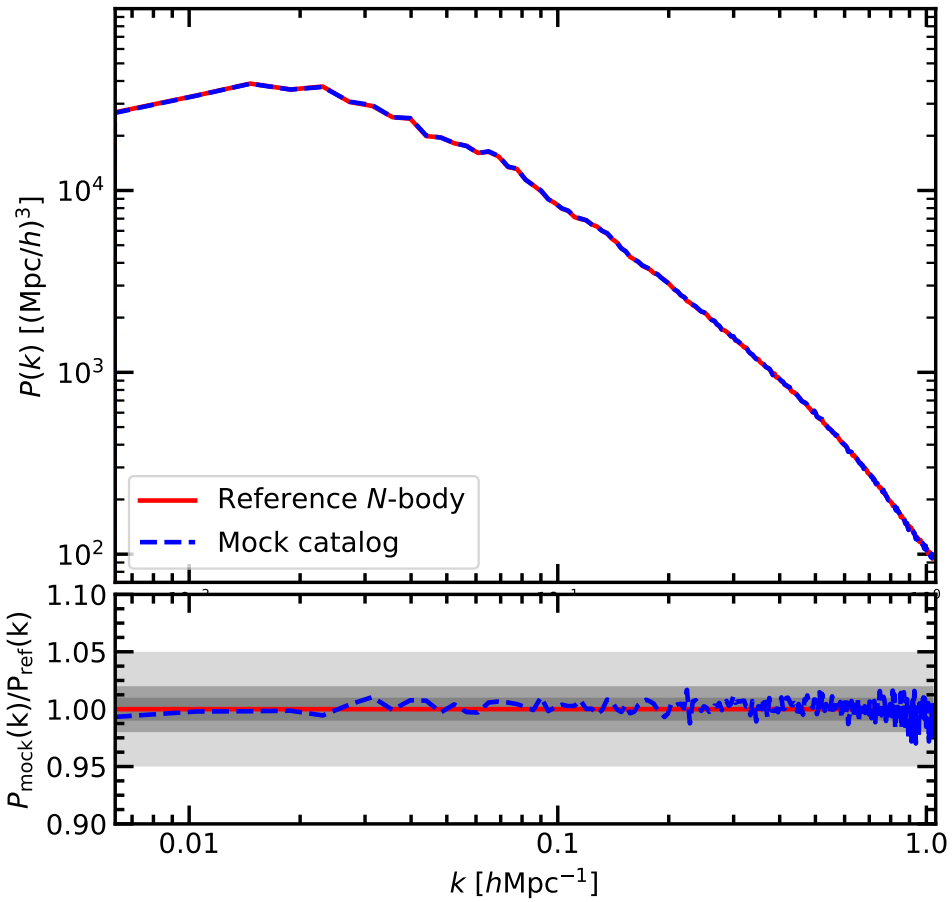

Figure 3: The plot shows an example of how the BAM method is able to produce halo mock catalogs with a very accurate power spectrum. The red solid line and blue dashed line represent respectively the $P(k)$ of the reference N-body simulation and the halo mock catalog produced by BAM. The upper panel shows the power spectra, the lower panel shows the ratio between mock and reference $P(k)$. The gray shaded areas represent respectively, from the darker to the lighter, $1\%, 2\%, 5\%$ residuals. It turns out that on large scales and intermediate scales the mock power spectrum is well-converged. On smaller scales, instead, the scatter is larger. The average residuals are $R < 1\%$.

assignment schemes and the main algorithms we make use are provided in Appendix A.

The strategy of BAM is organized in two steps:

- the extraction and learning of the bias, during which the code measures the bias from the accurate DMHF and DMDF of the reference simulation and then calibrates the relation to fulfill the accuracy requirement on the power spectrum;
- the sampling of mock catalogs, based on the calibrated bias relation and the approximated DMDFs obtained with ALPT.

The details of this procedure are described in what follows.

1 BIAS EXTRACTION

In this section we present how the extraction of the bias is performed.

As mentioned, this step needs a DMDF and a DMHF of a N-body simulation, from which the bias is extracted. Having one reference simulation is indeed one fundamental prerequisite to make BAM work.

1.1 Bias measurement

As anticipated in the previous sections, the bias is in this context interpreted in a statistical stochastic sense, inspired by Dekel & Lahav (1999). The number of halo is thought to be drawn from a probability distribution conditional to a set of properties of the underlying DMDF:

$$N_{\rm h} \backsim P(N_{\rm h}|\,\delta;\,\Delta V)_\lambda \quad , \tag{43}$$

where $N_{\rm h}$ is the number of halos in the cell λ, δ denotes the value of the dark matter density field contained in the cell and ΔV is the cosmic volume embedded in the cell.

Concretely, the measure of the bias is performed by recording cell by cell the values of the DM density and the halo number counts. Once this has been done, the code builds a 2-dimensional histogram having on one axis the dark matter density and on the other axis the number density (or equivalently, number counts) of halos in a cell. In other words, the bias is a matrix, whose pixels (its elements) represent the discretized values of the probability distribution, equipped with proper ranges of values along the axes. Concretely, the code assigns each cells to the pixel it belongs to, according to the definition of the quantities (number of halos and dark matter in this case) and the binning, and finally normalizes the probability distribution.

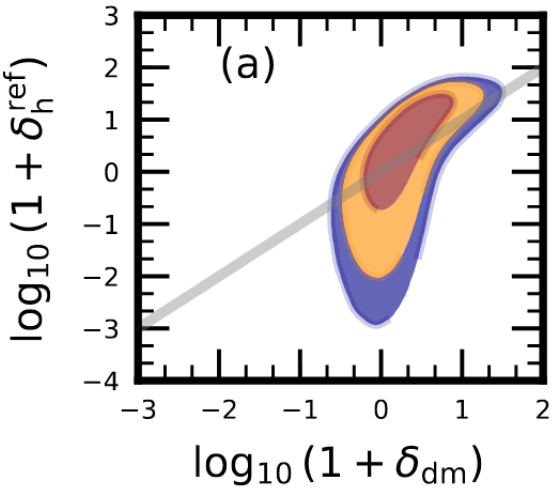

Figure 4: An example of the bias represented as joint probability distribution, measured from the dark matter density and dark matter halo fields of the MINERVA simulation. Contours denote the regions containing 68% (red), 98% (yellow) and 99% (purple) of the total number of cells. Taken by Balaguera-Antolínez et al. (2019).

Figure 4 shows an example of bias represented as joint probability distribution.

In the simplest approach depicted so far, the bias has been characterized only with DM density, which is by definition a local property. However, as discussed in chapter 2, the bias is in general regarded to be a complex non-linear functional of the dark matter field, which might include also non-local dependencies.

Therefore, we consider a more general definition of bias:

$$N_{\rm h} \backsim P(N_{\rm h}\,|\,\delta, {\rm X}, {\rm Y}, \ldots; \Delta V)_\lambda \quad , \tag{44}$$

where X and Y are the additional dependencies we want to account for. From a practical point of view, such generalization is achieved in the code by highering the dimensionality of the bias. In other words, if we want to include N dependencies (of which one is always DM density) in the bias, the bias will no longer be a matrix, but a $N+1-$rank tensor (N characterizing quantities + halos). Its element of position $\{x_1, \ldots, x_{n+1}\}$ represents the probability of finding a cell with the physical properties of such $N+1-$dimensional bin. The bias tensor is constructed as before: the code assigns the cells to the pixels they belong to and then normalizes the distribution.

At this point it is worth pointing out two facts which make the numerical construction of the bias not as trivial as one could think.

The 1-dimensional distribution of an arbitrary property included in the characterization of the bias is represented with a certain number of bins,

where in general the larger the number of bins, the higher the resolution and the accuracy we achieve in its representation. Binning the distribution of a quantity with, say, M bins means multiplying by M the number of bins used for the other properties. Therefore, if one describes $N+1$ properties (N properties + halo number counts) with M bins each, then the total number of bins used in the histogram is

$$N_{\text{bins}} = M^{N+1} \quad . \tag{45}$$

This makes it clear that the number of elements of the tensor grows rapidly with the number of properties added in the bias, and so does the memory requirement. Thus, if one needs to run BAM with many characterizing physical quantities, it becomes necessary either to use a low number of bins for each quantity or to use more performing super-computers. Anyway, this is not the case in the most common application of BAM.

The other important non-trivial numerical aspect lies in the way one does the binning itself. The binning procedure is in principle arbitrary, since one is free to choose the number of the adopted bins. However, some physical information is encoded in the 1-dimensional distribution of the considered characterizing physical quantity. Therefore, it is important to bin such quantity in a way which extracts the majority of the information contained, without adopting too many bins and thus make the memory requirements hard to be met. In general, one expects that highering arbitrarily the number of bins leads to a gain of the information encoded in the distribution of the considered quantity. However, at some point a saturation is reached and choosing a larger number of bins no longer results in a gain of information. This will be deepened in chapter 4, where an improved treatment of the variables included in the bias is presented.

1.2 First halo mock sampling and bias calibration

Once the code has learnt the bias relation from the reference simulation, it performs the first sampling of a halo mock.

Concretely, this is done as follows. The code considers each cell of the DMDF and sections the histogram representing the bias in correspondence of the value of DM density and all the other quantities used to characterize the bias. The result of such sectioning is the 1-dimensional probability distribution of finding certain N_{H} halo number counts in that cell. In other words, known all the values of the characterizing properties in the considered cell, we plug them in the general bias tensor and look at the resulting probability distribution of N_{H}. We can understand this operation in a clearer way by looking at Figure 4. What the code does in practice is slicing the bias with a virtual vertical line in correspondence of the value of DM density in the considered cell. The result of this operation is the

aforementioned 1-dimensional probability distribution $P(N_{\rm h})$. Eventually, the bias may be multi-dimensional, but this does not change the procedure.

At this point, a new DMHF (the future mock catalog) is initialized and each cell is assigned halo number counts with a random sampling, according to the PDF characterizing the homologous cell in the reference DMHF. This is done in such a way that the total number of assigned halos replicates the total number of halos of the reference DMHF.

In principle the code has in this way produced the first mock DMHF. However, the power spectrum of the mock DMHF in general is likely to deviate by $R \gg 1\%$ from the reference DMHF power spectrum. One might expect that the bias relation the code learns from the reference simulation should allow to sample mock catalogs with accurate power spectrum, since the bias is extracted from the full N-body calculation. However, a number of facts must be considered. First of all, the model of bias we are considering may not be complete from a physical point of view, or even partially incorrect. Indeed, there might be crucial additional dependencies which have not been included in our description, or on the contrary we may have included dependencies which does not carry physical information on the bias and act hence as noise. Then, the DM field has been interpolated on a mesh, which is in general an arbitrary way to represent a field. Moreover, in some cases the initial dark matter field has been down-sampled to a lower resolution (as happens e.g. with the application of `BAM` to the `MINERVA` simulation, see chapter 4). Finally, the halo field is extracted with a Friends-of-Friends halo-finder algorithm, which identifies halos as set of particles located within a certain linking length, which is however a bit arbitrary. Therefore, the bias relation may be affected by a lack of accuracy in this sense and produce, as indeed happens, an inaccurate mock catalog.

In order to overcome this issue and fulfill the convergence requirement, the code adopts an iterative process, to calibrate the bias relation in such a way to be able to perform the sampling of accurate mock catalogs. The production of the first mock DMHF is regarded as Iteration 0.

At this point it might be useful doing a clarification, in order not to confuse the reader. The mock catalog produced using the calibrated bias relation and the *same* dark matter field used for the calibration will be in general very similar to the reference DMHF. Precisely, let us suppose we were hypothetically able to completely model the halo bias numerically and physically at all scales. Then, once the bias relation is calibrated, the halo mock catalog produced by `BAM` would consist, beside random noise, in a reproduction of the reference DMHF. In other words, since we are using the same DM field, we will be sampling halo mock catalogs of the same realization. This is an important fact to be kept in mind. The studies and the findings presented in chapters 4 and 5 are indeed based on the idea of assessing the accuracy of different bias models on certain scales by analyzing how accurately a certain model allows to replicate the summary

statistics of the reference simulation. Therefore, we will sometimes refer to this procedure as *reproduction* of a field, instead of explicitly talking of *mock sampling*. The sampling of halo mock catalogs of different realizations will be discussed in section 2.

As mentioned previously, after the first mock sampling `BAM` does not achieve the target accuracy in the power spectrum, thus the code forces the power spectrum to convergence artificially. To this end, an iterative process is adopted. Such iterative procedure is structured as follows (the steps are carefully described in Balaguera-Antolínez et al. (2018)):

- after Iteration $i = 0$, a new DMHF is obtained. At this point, the code measures its power spectrum $P_{0j} = P_{i=0}(k_j)$, where k_j denotes the $j-$th spherical shell in Fourier space. The code also defines an isotropic kernel $\mathcal{K}_{ij}$ in Fourier space and initializes it to 1, i.e. $\mathcal{K}_{0j} = 1\,\forall j$;

- here Iteration $i = 1$ starts. A *bias transfer function* (BTF hereafter) is defined as $\mathcal{T}_{1j} = P_{\rm ref,j}/P_{0,j}$ and is assigned to the kernel, i.e. $\mathcal{K}_{ij} = \mathcal{T}_{1j}$. The values of the BTF for each k_j are selected with a Metropolis-Hastings algorithm. A transition probability $\min(1, \exp(\mathcal{H}^2_{0,j} - \mathcal{H}^2_{1,j}))$ is computed, where $\mathcal{H}_{ij} = (P_{\rm ref,j} - P_{i,j})/\sigma_j$ and σ_j is the Gaussian variance associated to the reference power spectrum. If the Metropolis-Hastings algorithm does not accept a $\mathcal{T}_{ij}$, `BAM` assigns to the kernel the value computed at the previous step $\mathcal{K}_{i-1,j}$;

- the DMDF is now convolved with the kernel and this operation generates a new DMDF. The code then measures again the bias relation between the new DMDF and the reference DMHF and uses it to do the sampling of a new mock DMHF. Then its power spectrum $P_{1,j}$ is measured, the kernel and the BTF are updated and the iterative process restarts;

- at Iteration i the code calculates the BTF as $\mathcal{T}_{ij} = P_{\rm ref,j}/P_{i-1,j}$ and the kernel is updated as $\mathcal{K}_{ij} = \mathcal{T}_{1j} \times \cdots \times \mathcal{T}_{i-1,j} \times \mathcal{T}_{ij}$ by using the Metropolis-Hastings algorithm as described above;

- the iterative process goes on until the convergence is achieved. The residuals are averaged over all k-bins in Fourier space and typically one observes $R \lesssim 1\%$. However, it is interesting looking at how the $P(k)$ converges on different scales. In the case of halo catalogs, for instance, the $P(k)$ tends to be well-converged at intermediate scales, while can present a larger deviation at large scales, affected by the cosmic variance, and at small scales, characterized by non-linearities and by the aliasing in the measure of $P(k)$ (see Jing, 2005).

In general, few tens of iterations $(20 - 40)$, depending on the number of quantities used to characterize the bias, are enough to achieve the target

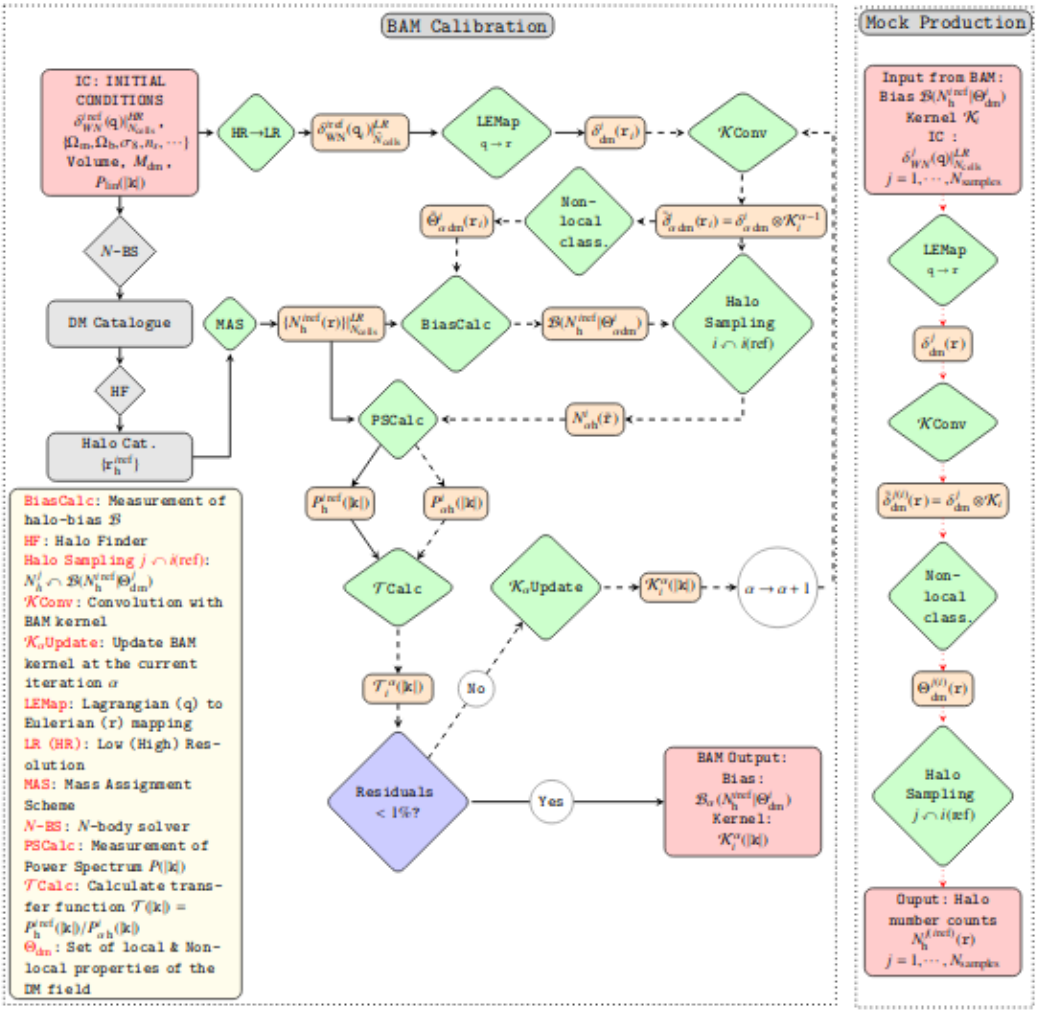

Figure 5: This flowchart illustrates the overall ensemble of operations performed within BAM.

accuracy. Figure 5 illustrates the whole set of operations performed within BAM. Figure 6 shows an example of how the mock power spectrum converges to the reference one along the iterations.

One crucial point regards the 3-points statistics, studied throughout this work by means of the reduced bispectrum $Q(\theta_{12}, k_1, k_2)$, introduced in chapter 2. While in fact the power spectrum of the mock halo catalog is constrained to converge to the reference power spectrum by construction, the code does not artificially alter the 3-points statistics. This means that the bispectrum of the mock catalog is exclusively determined by the bias. As said, when we use the same realization of the dark matter field, the produced mock replicates the reference halo field, with accuracy depending on the adopted bias dependencies. Thus, the better the bispectrum of the mock fits the bispectrum of the reference DMHF (mock and reference bispectrum, hereafter) on some cosmological scales, the more complete the bias characterization at those scales is thought to be. The motivation for this is that the practical effect of including additional dependencies in the bias, besides local DM density, is to introduce constraints on the region of the parameter space in which a cell resides. Concretely, this means restricting the 1-dimensional probability distribution (before normalization) to a lower number of cells, i.e. only the cells which are within the considered multi-dimensional bin. Consequently, because the halo random sampling is performed according to such probability distribution, if the constraints introduced by the additional properties physically characterize the bias in a correct way, then the halo number counts assignment will be more accurate. On the contrary, it may happen that accounting for wrong dependencies in the bias restricts the parameter space to an incorrect region and therefore the code loses accuracy in the assignment of number counts. Therefore, the 3-point statistics optimally assesses the accuracy of a halo mock, with respect to the reference halo field. For this reason, bispectra studies represent one of the fundamental aspects investigated throughout this work.

The reference BAM papers show that the reduced mock bispectrum $Q(\theta_{12}, k_1, k_2)$, for large-scale configurations such as $k_1 = 0.1$, $k_2 = 0.2\,h\mathrm{Mpc}^{-1}$, features a systematic deviation of $10 - 30\%$ with respect to the reference bispectrum, for different considered bias models. The expectation is that the BAM method can be used to investigate which are the crucial dependencies that fully characterize the bias on a certain scale. This aspect is studied in details in Kitaura et al. (2020)[1] and presented in chapter 4.

[1] The author of the Master Thesis is also co-author of this paper.

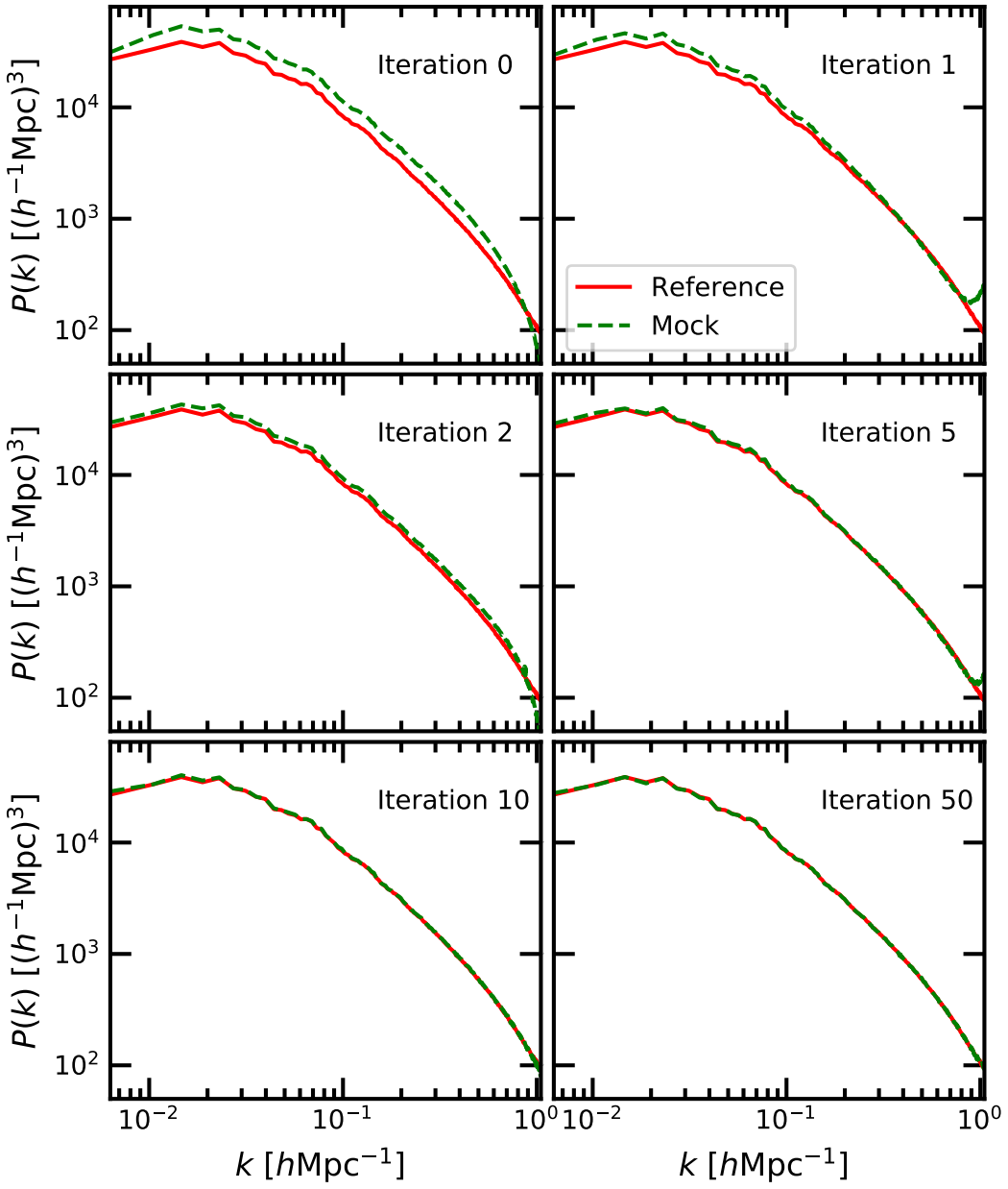

Figure 6: This plot illustrates how, thanks to the iterative process employed by `BAM`, the mock power spectrum converges to the reference power spectrum along the iterations.

2 HALO MOCK CATALOGS SAMPLING

Even though the sampling of halo mock catalogs (of different realizations) goes beyond the scope of this work, it represents the ultimate aim of all the calibration and bias characterization studies. Therefore, we briefly present here the mock-sampling procedure.

The bias calibration results in a relation which, except for arguments of cosmic variance, is assumed to be valid in all the regions of the Universe at the same redshift. Therefore, if one has available different realizations of the dark matter field, with the same resolution and redshift of the reference DM field, they can be used to sample halo mock catalogs by mapping the halo distribution through the bias.

The procedure is identical to the one adopted for the mock-samplings performed in the calibration and we sketch it as follows:

- the code considers the bias relation, calibrated on the i-th realization, and the j-th realization of the DM field;
- each cell of the dark matter mesh field is assigned to the multi-dimensional bin to which it belongs, according to the generalization of the bias as multi-variate probability distribution of dimensionality $N + 1$, given N characterizing quantities;
- the bias is sliced in correspondence of the bin of the considered cell, getting as a result the 1-dimensional probability distribution $P(N_{\rm h})$;
- halo number counts are randomly sampled in each cell according to $P(N_{\rm h})$.

The overall distribution of halos will in the end trace the spatial distribution of the dark matter.

As mentioned, the dark matter fields can be obtained in the scope of `BAM` by evolving ICs of different realizations with ALPT. This in principle produces approximated dark matter fields, which are not as accurate as the dark matter field obtained with the full N-body calculation. However, the accuracy is high enough to successfully perform the sampling of accurate halo mock catalogs. The accuracy is eventually improved with the use of phase-space mapping (Abel et al., 2012; Hahn et al., 2013).

3 BAM APPLIED TO CONTINUOUS TRACERS

In this section we introduce how `BAM` is applied to continuous Large Scale Structure tracers.

As mentioned, `BAM` has been designed to sample mock catalogs of discrete tracers of the underlying dark matter distribution, such as dark matter halos

and galaxies. In other words, it has been built to sample number counts in cells. However, the method has been extended to cope with continuous tracers, such as e.g. the main constituents of intergalactic medium.

This is an aspect of paramount importance. If indeed the method is shown to perform with the same high accuracy achieved for dark matter halos, this may potentially allow to sample accurate mock catalogs of properties of the intergalactic medium without the need of running full hydrodynamical simulations of large cosmological volumes. Such simulations are indeed particularly expensive due to the need of solving both gravity and the hydrodynamics. As in the case of halo mock catalogs, the time-scale for mock production is expected to be of the order of tens of seconds, to be compared with hundreds/thousands of hours elapsed in large volume hydrodynamical simulations.

The extension of BAM to continuous quantities requires just minimum modifications. The idea is the following:

- first, the considered continuous quantity is properly binned. Let us consider a quantity x, spanning the interval $[x_{\rm min}, x_{\rm max}]$, to be described in a histogram with n bins, such that each bin has width $\Delta x = |x_{\rm max} - x_{\rm min}|/n$. A precise value $\tilde{x}$ of x is placed in the bin with edges $\{x_{\rm i}, x_{\rm i+1}\}$, such that $x_{\rm i} < \tilde{x} < x_{\rm i+1}$;

- an integer number is assigned to each bin, in such a way that different bins are assigned different numbers. The assigned integer number represents *fictious number counts* and is computed as:

$$n_{\rm int} = {\rm floor}\left(\frac{\tilde{x} - x_{\rm min}}{n}\right) \quad .$$

  In practice, the integer number assigned to a bin corresponds to its index;

- BAM seeks the bin to which a cell belongs to and performs the same assignment described in the previous sections. Notice that at this point the bins of the considered property are labelled with integer numbers and the assignment becomes identical to the one of a discrete quantity. Thus, the code assigns to a cell number counts, chosen randomly according to the probability distribution $P(N_{\rm h})$;

- once number counts have been assigned, the code re-transforms back the assigned integer number to a real value $\tilde{x}$ of the continuous field x by means of the following transformation:

$$\tilde{x} = x_{\rm min} + (n_{\rm int} + 0.5)\Delta x \quad .$$

  Notice that $\tilde{x}$ corresponds to the middle value of the bin to which it has been assigned.

Finally, a mock of the considered continuous tracer field is produced. The basic idea we adopt is hence to make it mimic a discrete field and perform a number counts assignment.

In principle, one might expect this procedure to be not as accurate as the one employed for dark matter halos, due to the intrinsic continuous nature of the considered quantity. However, in chapter 5 we will show that BAM is able to replicate very accurately ionized intergalactic gas density, neutral hydrogen number density, temperature and optical depth fields. From a qualitative point of view, the mock catalogs turn out to be visually almost indistinguishable from the reference simulation. Quantitatively, the mock power spectrum is found to achieve percent accuracy (in most of the cases) and the mock bispectrum is found to be compatible within error bars with the reference bispectrum.

Of course, to perform cosmological studies one needs a reference hydrodynamical simulation with high resolution and large volume, the latter of the order $V \gtrsim 100\,h^{-3}\mathrm{Mpc}^3$. This is one non-trivial aspect to be considered, since very few research groups in the World run such simulations, which are computationally super expensive.

4 NON-LOCAL HALO BIAS STUDY WITH BAM

In this chapter we present how BAM is applied to extract the halo bias relation, introduced in chapter 2, and to investigate the non-local contribution provided by the dark matter cosmic web.

In chapter 3 we introduced the BAM method and described how it generates mock catalogs of a given class of tracers of the underlying dark matter distribution. In particular, let us remind that BAM samples mock catalogs with very accurate 2-point statistics by construction. However, the 3-point statistics of the mock catalogs is not artificially forced to converge to the one of the reference simulation. Improvements in the reproduction of the target 3-point statistics are due to stronger constraints on the spatial distribution of the tracers introduced in the bias relation. Therefore, we regard the 3-point statistics is an optimal indicator of the accuracy of the mapping we are adopting.

This chapter is organized as follows. In section 1 we present the details of the MINERVA simulation, the reference N-body cosmological simulation from which BAM learns the bias relation. Section 2 explains how we apply BAM to study the halo bias and which bias models we intend to test. In section 3 we present the quantitative results of our analysis and show that a proper modelling of the dark matter cosmic web (see chapter 2) is sufficient to completely characterize the halo bias on large scales (e.g. $k_1 = 0.1\,h\mathrm{Mpc}^{-1}$, $k_2 = 0.2\,h\mathrm{Mpc}^{-1}$ in Fourier space). Finally, in section 4 we give some outlooks on possible future extensions of this work.

The main findings reported in this chapter are presented in details in Kitaura et al. (2020).

1 THE MINERVA SIMULATION

The reference simulation employed here is the MINERVA (Grieb et al., 2016), which consists in a suite of 300 N-body cosmological simulations. It is run using the code Gadget2 (Springel, 2005), with $N = 1000^3$ dark matter particles in a volume $V = 1500\,h^{-3}\mathrm{Mpc}^3$. Such volume is large enough to analyze anisotropic galaxy clustering with a precision comparable to present-day galaxy redshift surveys. The simulations employ periodic boundary conditions and are started at redshift $z = 63$, with initial conditions generated with 2LPT (second-order Lagrangian Perturbation Theory). The input linear power spectrum for the initial conditions was calculated using Camb (Lewis et al., 2000). The snapshots are stored for redshift $0 < z < 2$ and the final

halo catalogs are generated with the SubFind code (Springel et al., 2001). The minimum resolved mass of halos is $m_{\rm min} = 2.67 \times 10^{12}\, h^{-1}{\rm M}_{\odot}$.

In our study, we make use of the output (namely, the dark matter density field and the corresponding halo field of one of the 300 realizations) at redshift $z = 1$, interpolated on a mesh downsampled to a resolution of $N = 500^3$ cells.

2 LOCAL AND NON-LOCAL HALO BIAS

In Kitaura et al. (2020) we investigate an effective halo bias model at the field level (see Schmittfull et al., 2019), i.e. at the level of realization. In other words, we start from a realization of the dark matter density field traced by a halo field, and investigate the halo bias by studying the summary statistics of the halo field. This approach differs from works which adopt the perspective of assuming a certain Perturbation Theory bias expansion and compute predictions which works well for the summary statistics (see Desjacques et al., 2018, for a recent review), i.e. halo power spectrum, cross-power spectrum between halo and matter field and higher-order correlation functions. Indeed, while studying the bias by reproducing the halo field of a given realization automatically replicates also the summary statistics, the other way around is in general not true.

We employ here the following model, obtained from equation 35 by restricting the bias to only some of the dependencies:

$$\delta_{\rm h}(\vec{r}) \backsim P\left(\delta_{\rm h}(\vec{r})|\, \nabla^2\phi(\vec{r}),\, \partial_i\partial_j\phi(\vec{r})\right) = P\left(\delta_{\rm h}|\, \delta,\, \mathcal{T}\right)\,. \tag{46}$$

i.e. we express the probability of finding a certain halo overdensity $\delta_{\rm h}(\vec{r})$ as a function of the local dark matter overdensity δ and the symmetric tidal field tensor $\mathcal{T} = \partial_i\partial_j\phi(\vec{r})$.
This relation is inspired by equation 38, but is not truncated at any order. It includes the infinite non-linear expansion of δ and the tidal anisotropy terms to higher orders and corresponds to Resummed Perturbation Theory (see Crocce & Scoccimarro, 2006; Elia et al., 2011; Matarrese & Pietroni, 2007).

As done in chapter 2, we can subdivide such dependencies into local and non-local. By definition, the evolved non-linear overdensity δ and its higher orders represent local halo-bias terms. The terms which are built from the gravitational tidal field tensor are instead non-local and long-range, since they account for the underlying geometry of the gravitational potential, which depends on the (anisotropic) distribution of dark matter at long-range. In our study we neglect the contribution of peculiar velocities and of eventual short-range non-local bias terms. The two are to be investigated in future works.

A model of local bias, depending only on dark matter density, has been already investigated in the reference BAM papers. Moreover, such works have included in the bias the dependencies on the dark matter cosmic web, in the T-web definition, and on the mass spectrum of the large-scale collapsing regions (also called super-knots). This attempt led to a significant improvement of the 3-point statistics of the mock catalogs and such result has been regarded as a possible signature of non-local bias.

Motivated by these results, we have tested more complex models including local and non-local dependencies, to find the optimal mapping of halos on the dark matter field and to assess the impact of different non-local terms. The reliability of the models has been quantitatively assessed by analyzing how accurately they allow to replicate the 3-point statistics of the reference simulation on large scales. We refer to chapter 2 for the definitions regarding the tidal field tensor $\mathcal{T}$, its eigenvalues $\{\lambda_i\}$ and its invariants $\{I_n\}$, the bias perturbative expansion and the I-web and T-web definitions of cosmic web.

Following the notation of Kitaura et al. (2020), we start by expressing a certain variable q as a functional dependence on some variables, regarded as generating functions $\{\eta_1, \eta_2, \dots\}$:

$$q = F(\eta_1, \eta_2, \dots) \quad . \tag{47}$$

We then assume that the BAM method is able to express q as a linear combinations of the non-linear expansion of each of the generating functions:

$$q = c_1 F_1(\eta_1) + c_2 F_2(\eta_2) + \dots = c_1 F_1(a_1\eta_1 + a_2\eta_1^2 + \dots) + c_2 F_2(b_1\eta_2 + b_2\eta_2^2 + \dots) + \dots \quad . \tag{48}$$

In other words, if we introduce in the bias description a number of generating functions, which in this concrete case are explicit non-linear products of the $\{\lambda_i\}$, BAM is able to linearly combine their non-linear expansions. In order to test the validity of this statement, we build an *ad hoc* experiment. We consider separately the following models:

- λ-web: $\{\lambda_1, \lambda_2, \lambda_3\}$;
- PT-web: $\{\delta, I_4, I_5\}$, where $I_4 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$ and $I_5 = \lambda_1^3 + \lambda_2^3 + \lambda_3^3$.

If BAM is really able to perform the linear combination of non-linear expansions of the generating functions described above, then it should treat the two models as equivalent. This claim is confirmed by the results of the BAM runs. The right panel of Figure 7 shows the reduced bispectrum of the mock catalogs obtained respectively with λ-web and PT-web, at configuration $k_1 = 0.1$, $k_2 = 0.2\,h\mathrm{Mpc}^{-1}$ in Fourier space. Since the two bispectra are in perfect agreement and compatible within error bars, we regard BAM handles

such models in an equivalent manner. This means that BAM is able to expand the λ_i into higher-order powers of each of them.

Then, we want to understand which is the minimum number of variables required to constrain q and be able to reproduce it by including such generating functions in the bias description in BAM. In our case, q is the halo overdensity and the generating functions are:

- local DM overdensity δ and higher-order powers δ^2, δ^3, ... The latter are automatically accounted for by the non-linear expansion performed by BAM that we have just described;
- non-local terms consisting in the principal invariants of the tidal field tensor: $\{I_1,\ I_2,\ I_3,\ I_4,\ I_5\}$.

Detailed calculations in Kitaura et al. (2020) show that we can identify the main invariants I_4 and I_5 with the second and third order non-local bias terms s^2 and s^3 respectively (up to some powers of δ, due to the different definitions of $\mathcal{T}_{ij}$ and s_{ij}). Moreover, it has been shown that a model such as $\{I_1,\ I_2,\ I_3,\ I_4,\ I_5\}$ (dubbed I-web-c), which include all the invariants of $\mathcal{T}$, is sufficient to properly model the terms s^2, s^3, $s^2\delta$, the ellipticity e and the prolateness p. That is, I-web-c fully constrains all the terms of the perturbative expansion presented in McDonald & Roy (2009) up to 3rd order, which are the terms that dominate the statistics within the scales we are considering.

Therefore, we concretely consider the following models (following again the notation of Kitaura et al. (2020); $\tilde{x}$ indicates a partial modellization of a variable):

- Local DM δ:
- δ+T-web: $\{\delta,\ t_{\rm cw}$={knot, filament, sheet, void}$\} \sim \{\delta(I_1), \widetilde{s^2}(\tilde{\alpha}), \widetilde{s^3}, \tilde{e}, \tilde{p}\}$,
- PT-web-q: $\{\delta, I_4\}$;
- PT-web: $\{\delta, I_4, I_5\}$;
- λ-web: $\{\delta, \lambda_1, \lambda_2, \lambda_3\} \sim \{\delta, I_4, I_5, e, p\}$;
- I-web: $\{\delta, I_2, I_3\} \sim \{\delta, s^2(\alpha), \widetilde{s^3}, \widetilde{\delta_{\rm L} s^2}, e, p\}$;
- I-web-c: $\{\delta, I_2, I_3, I_1 I_2\} \sim \{\delta, I_2, I_3, I_4, I_5\} \sim \{\delta, s^2(\alpha), s^3, \delta_{\rm L} s^2, e, p\}$.

Among these, local DM δ is the bias local bias model. δ+T-web is the long-range non-local model which accounts for the cosmic web dependence defined as in Hahn et al. (2007), while I-web and I-web-c are the models based on the I-web re-definition of cosmic web, the latter modelling more completely the 3rd order of the bias expansion. λ-web and PT-web have

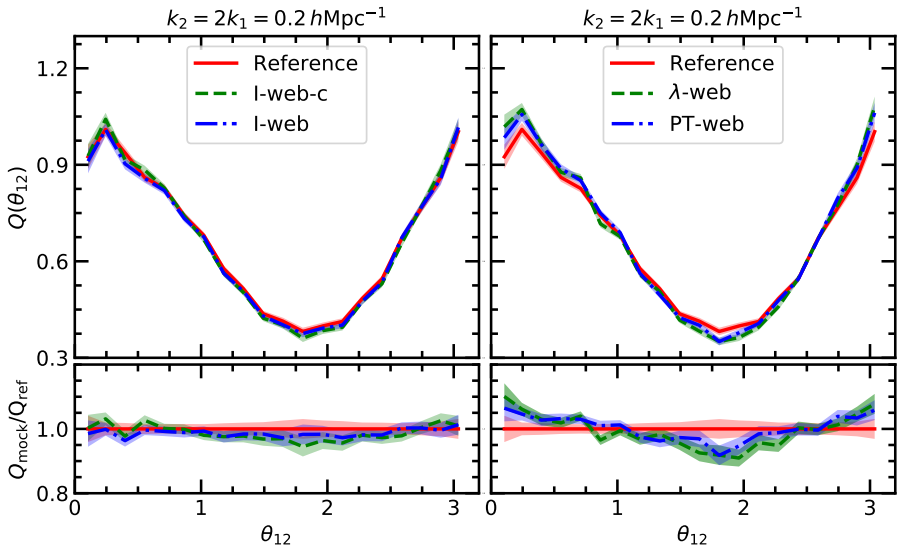

Figure 7: In the upper panels, reduced bispectra; in the lower panels ratios $Q_{\mathrm{mock}}(\theta_{12})/Q\mathrm{ref}(\theta_{12})$ at configuration $k_1 = 0.1, k_2 = 0.2h\mathrm{Mpc}^{-1}$ in Fourier space. In the panels on the left, a comparison between reference bispectrum (red solid line) and mock bispectra, produced respectively with the I-web (blue dash-dotted line) and I-web-c (green dashed line). On the right, the same plots but with λ-web (green dashed line) and PT-web (blue dash-dotted line). The results tell us that I-web-c does not add a significant amount of information with respect to I-web and that the λ-web and PT-web produce in BAM equivalent results, at the investigated scales.

already been introduced previously and are though to be equivalent. PT-web-q is a model constructed with the same philosophy of PT-web, i.e. including terms which do not mix up different eigenvalues, but is restricted to 2nd order. The unique bias formulations among the ones considered which provide a complete modellization of the bias expansion up to 3rd order are I-web-c and I-web. The latter actually models only partially some terms at 3rd order, which are however shown not to be crucial (later in this section). Therefore, our numerical analysis consists of a number of numerical tests aimed at probing these considerations and showing that I-web and I-web-c are the models which allow to replicate better the 3-point statistics of the reference simulation in the mock catalogs.

Before proceeding with the numerical analysis, let us try to understand the operative difference between I-web and I-web-c. To this end, we look at the resulting mock bispectra, shown in the left panels of Figure 7. The upper panel shows reduced bispectra $Q(\theta_{12}, k_1, k_2)$, while the lower panel shows ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$ between mock and reference bispectra. Even though I-web-c in principle fully constrains all the 3rd order terms while I-web models some of them only partially, the mock bispectra produced by

the two are in perfect agreement. This fact suggests that the terms which I-web models only partially are not crucial at these large scales. Moreover, some analytical arguments presented in Kitaura et al. (2020) show that the term $I_1 I_2$ is related to I_2 and I_3 through the equation

$$I_1 I_2 = 3I_3 + \sum_{i \neq j} \lambda_i \lambda_j^2 \quad . \tag{49}$$

This means that when the eigenvalues are similar, which happens e.g. in knots and voids, $I_1 I_2 \sim 9 I_3$ and hence, including explicitly the generating function $I_1 I_2$ (which is the one that diversify I-web from I-web-c) does not add much information in the bias.

Therefore, we will assume from now on the I-web model is representative also of a more complete model, I-web-c, and use it throughout the rest of this work. This choice allows to save one variable in the numerical modelling and thus to make the computational implementation faster and cheaper in terms of memory requirements.

3 RESULTS

In this section we present the results of the numerical tests involving the halo bias models presented in the previous section. We will find that I-web allows to improve the precision of the 2-point statistics of the mock catalogs and reproduce the reference 3-point statistics with unprecedent accuracy.

3.1 An improved treatment of variables

In order to account efficiently for certain generating functions in the halo bias, it is crucial to model their binning properly, in such a way to extract the most of the information encoded in such variables. In other words, one has to represent the probability distribution of a given variable with a number of bins which is suited to the specific analyzed quantity. This task, which may appear straightforward to be performed, is non-trivial in our context. In many cases, indeed, the variables are characterized by distributions with a very narrow central peak and very long tails. In a more mathematical language, one may say such distributions are characterized by a very large kurtosis. If indeed we compared the distributions with a Gaussian having the same variance, then the tails of the distribution would approach zero much more slowly than the Gaussian. As a result, when a linear binning is adopted, either one uses a huge amount of bins or gives up in properly resolving the central peak and the information encoded there.

One clear example of such distributions is provided by the invariants of the tidal tensor I_2 and I_3. Their distribution span respectively the intervals

$[-348.1, 5568.9]$ and $[-4175.7, 79633.1]$, but the $> 97\%$ of the cells are classifies within $[-1, 1]$ (in proper units). To overcome this problem we suggest to apply a non-linear transformation such as $y = x^{\alpha}$, i.e. a power-law, with $\alpha < 1$. The choice of the exponent is motivated by the fact that we need a transformation which relaxes the central peak and drives instead the very large positive (negative) values the closer possible to $y = 1$ ($y = -1$). In this case, we choose $\alpha = 1/9$, even though we have verified that the choice of a specific exponent does not lead to a better result than other exponents, provided that the effect of the transformation pursues the goal for which has been designed. Figure 8 shows the result of this operation. First and second rows show respectively the distributions of I_2 and I_3. The left panels show the original distributions, while the right panels show the transformed distributions. It turns out to be clear that their original distributions are not adequately represented, while the transformed distributions are.

If the transformation is not applied, the code sees the original variable, with its original distribution. When this happens and the distribution is poorly described, BAM acts as if it was not considering such variable, because it cannot infer useful constraints on the spatial distribution of the tracers from that generating function.

This transformation is crucial throughout this work. We apply it to many quantities, since many of them have distributions which share the features described above.

3.2 Qualitative assessment

We start our analysis with a visual inspection of the results.
Figure 9 shows in the upper panel a slice of the reference dark matter field, in blueish colours, with a slice of the reference halo field overplotted in red. The mid panel shows, in the same colours, the approximated dark matter field produced with ALPT and the halo mock catalog produced by BAM, once I-web is included in the bias. The lower panel shows the difference between the two. It turns out that top and mid panels are visually in good agreement, indistinguishable by eye. This is a good indication of the fact that BAM has accurately reproduced the reference DM halo field. By looking at residuals in the bottom panel, one notices that they resembles very closely the aspect of random noise and no systematic differences are detected. In fact, let us recall that there is always a stochastic noise component ϵ which is not modelled in BAM.
However, to be precise, we need to perform a quantitative analysis.

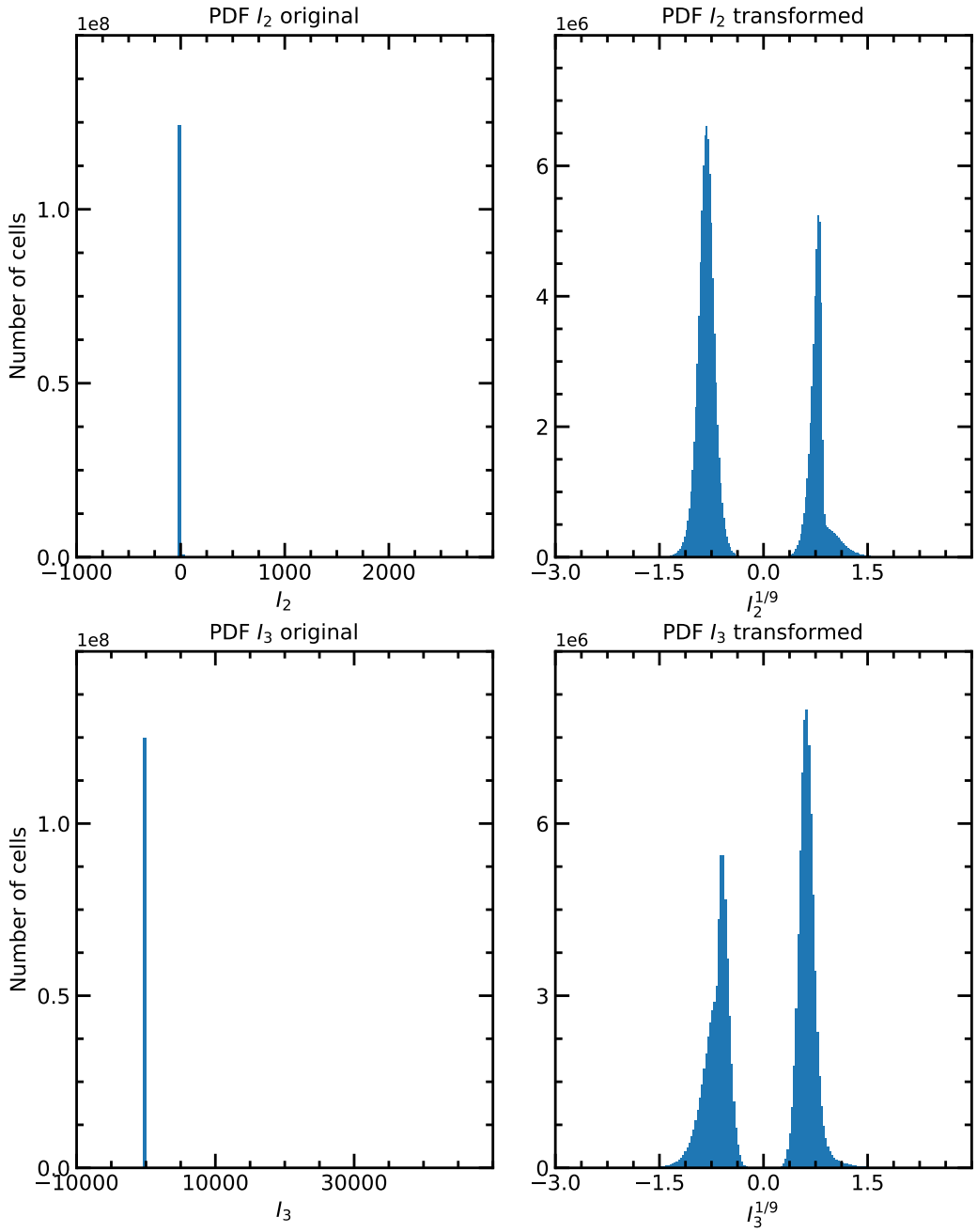

Figure 8: Comparison between original and transformed distributions of the generating functions to be included in `BAM`. First and second rows show respectively the distributions of I_2 and I_3. The left panels show the original distributions, while the right panels show the transformed distributions. It turns out that the original distributions do not describe the encoded information in a proper way, while the transformed distributions do.

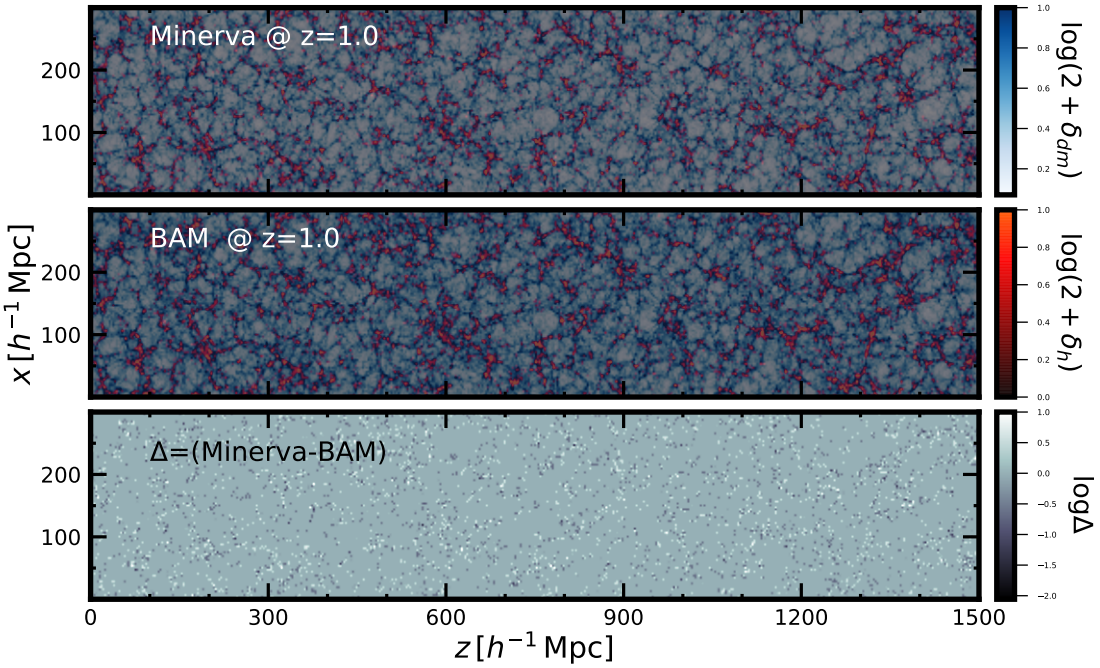

Figure 9: Comparison between slices of the reference simulation and the mock catalog created by BAM. In the upper panel a slice of the reference dark matter field, in blueish colours, with a slice of the reference halo field overplotted in red. The mid panel shows, in the same colours, the approximated dark matter field produced with ALPT and the halo mock catalog produced by BAM, once I-web is included in the bias. The lower panel shows the residuals.

Mock	$R(\%)$	$\sigma(\%)$
Local DM δ	0.74	0.94
T-web	0.75	0.84
PT-web	0.78	0.66
PT-web-q	0.73	0.97
λ-web	0.63	0.71
I-web	0.60	0.54

Table 1: Percentage residuals R and standard deviation σ of $P_{\rm mock}(k)/P_{\rm ref}(k)$ ratios for the halo mocks obtained with the considered bias models. R and σ are computed up to $k \sim 0.72h{\rm Mpc}^{-1}$, i.e. $k \sim 0.7\,k_{\rm nyq}$.

3.3 Quantitative assessment

In this section we describe the results BAM achieves in the reproduction of the reference summary statistics, with different bias models.

Summary statistics: power spectrum

Figure 10 and Figure 11 show in the upper panels the power spectra $P(k)$ of the reference DM halo field and of different mock catalogs, while the lower panels show the ratios $P_{\rm mock}(k)/P_{\rm ref}(k)$.

The upper panels clearly illustrate how BAM produces by construction very accurate power spectra. Indeed, all the mock $P(k)$ fit very accurately the reference $P(k)$.

The lower panels allow to draw more quantitative conclusions on the impact that different models have on the power spectrum. Even though the mock power spectra are all well-converged to the reference power spectrum, some clear improvements are achieved using I-web instead of other bias models. The power spectrum of the mock catalog produced with I-web is found to deviate from the reference $P(k)$ less than the other models on large scales, $0.006 < k < 0.03\,h{\rm Mpc}^{-1}$, which tend to be affected by cosmic variance. On the whole, it never exceeds 1% residuals (darker gray shaded area) and reduces significantly the scatter on small scales too.

On one hand, I-web leads to a significant improvement with respect to local DM density density and δ+T-web models. These two models had already been explored in Balaguera-Antolínez et al. (2018), Balaguera-Antolínez et al. (2019), Pellejero-Ibañez et al. (2020). On the other hand, it produces a mock power spectrum which is more precise than the ones obtained with models inspired by Perturbation Theory, namely PT-web and PT-web-q, and with λ-web. Table 1 shows average percentage residuals R and the standard deviation σ of ratios $P_{\rm mock}(k)/P_{\rm mock}(k)$, limited to spherical shells in Fourier space up to $k \sim 0.72h{\rm Mpc}^{-1}$, i.e. $k \sim 0.7\,k_{\rm nyq}$. I-web clearly achieves the best precision among the investigated models.

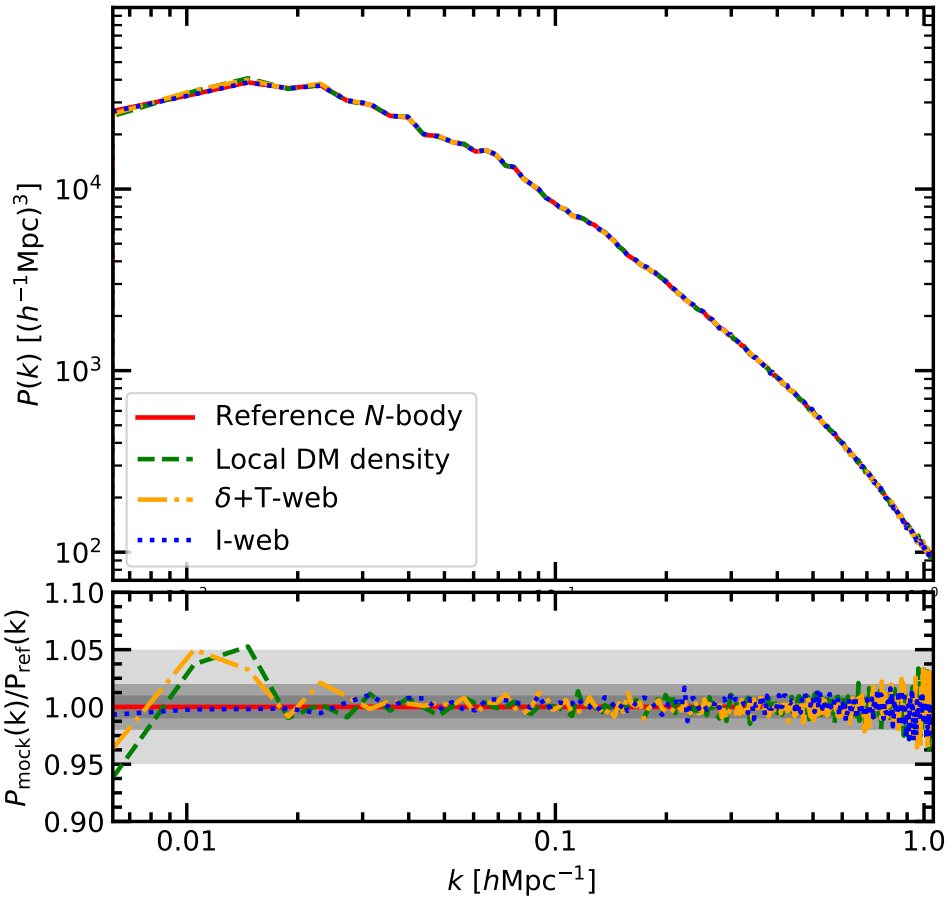

Figure 10: This plot shows the power spectrum of the reference DM halo field, represented by the red solid line, compared to power spectra of mock catalogs employing different bias prescriptions: local DM density (green dashed line), T-web (yellow dash-dotted line) and I-web (blue dotted line). The upper panel show the $P(k)$, while the lower panel shows the ratio $P_{\mathrm{mock}}(k)/P_{\mathrm{ref}}(k)$. Gray dashed areas stand respectively, from the darker to the lighter, for $1\%, 2\%, 5\%$ residuals.

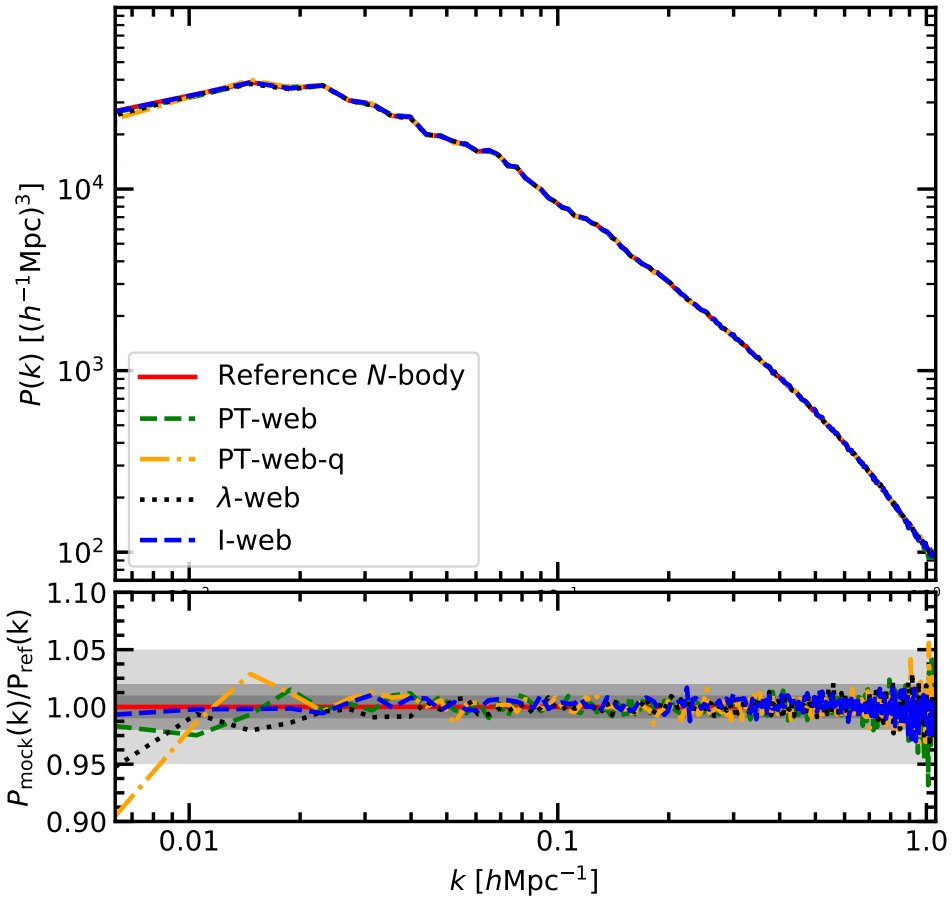

Figure 11: This plot shows the power spectrum of the reference DM halo field, represented by the red solid line, compared to power spectra of mock catalogs employing different bias prescriptions: PT-web (green dashed line), PT-web-q (yellow dash-dotted line), λ-web (black dotted line) and I-web (blue dashed line). The upper panel show the $P(k)$, while the lower panel shows the ratio $P_{\mathrm{mock}}(k)/P_{\mathrm{ref}}(k)$. Gray shaded areas stand respectively, from the darker to the lighter, for $1\%, 2\%, 5\%$ residuals.

Summary statistics: bispectrum

The second part of the quantitative analysis focuses on the 3-point statistics. Concretely, we investigate the reduced bispectrum and analyze which among the considered models allows to reproduce the bispectrum of the reference simulation more accurately in the mock catalogs. The analysis has been performed on large scales, $0.02\,h\mathrm{Mpc}^{-1} < k < 0.2\,h\mathrm{Mpc}^{-1}$, leaving smaller scales for future works.

The main findings are shown in Figure 12. In the first row, reduced bispectra $Q(\theta_{12}, k_1, k_2)$; in the second row, ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$ between mock and reference bispectra. The red solid line represents the bispectrum of the reference halo catalog, while green dotted line represents the mock bispectrum obtained with local DM δ, yellow dash-dotted line the one obtained with δ+T-web, the black dash-dotted line the one obtained with PT-web and the blue dashed line the one obtained with I-web. The third row shows the significance of each of the tested model with respect to the reference, computed as:

$$\sigma_{\mathrm{ref}}(\text{X-web}) = |\langle Q_{\text{X-web}}\rangle - \langle Q_{\mathrm{ref}}\rangle|/(\sqrt{\sigma^2_{\mathrm{ref}} + \sigma^2_{\text{web}}}) \tag{50}$$

The fourth row shows the significance $\sigma(\mathrm{X}, \mathrm{Y})$ of the X-web model with respect to the Y-web one, computed as

$$\sigma_{\mathrm{web}}(\mathrm{X}, \mathrm{Y}) = |\sigma_{\mathrm{ref}}(\text{X-web}) - \sigma_{\mathrm{ref}}(\text{Y-web})| \quad . \tag{51}$$

In first and second rows, error bars, represented as shaded areas, are calculated as standard deviations of bispectra of the 300 halo catalog realizations of the `MINERVA` simulation. Strictly speaking, such errors are valid only for the reference bispectrum. However, Balaguera-Antolínez et al. (2019) showed that the errors that one gets by averaging the bispectra of mock catalogs produced with `BAM` are comparable with the ones obtained by averaging the bispectra of the 300 `MINERVA` realizations. For this reason, the errors computed from the reference halo fields has been used for all the mocks too.

Different columns represent different configurations in Fourier space. The investigated cosmological scales are (from the left to the right): $k_1 = k_2 = 0.02\,h\mathrm{Mpc}^{-1}$, $k_1 = k_2 = 0.05\,h\mathrm{Mpc}^{-1}$, $k_1 = k_2 = 0.1\,h\mathrm{Mpc}^{-1}$ and $k_1 = 0.1$, $k_2 = 0.2\,h\mathrm{Mpc}^{-1}$.

At the largest scales, $k_1 = k_2 = 0.02\,h\mathrm{Mpc}^{-1}$, in the leftmost column, one notices that the noise is dominant because the triangular configurations are not many (due to the limited volume of the box), error bars are large and all the mock bispectra reproduce the reference well within error bars. At such scales, we cannot establish confidently which among the mock bispectra replicates better the reference bispectrum.

However, when one goes down to smaller scales, the noise is significantly reduced and differences in the mock bispectra produced with different bias models arise. In particular, let us focus on the smallest configuration we have considered, $k_1 = 0.1$, $k_2 = 0.2\,h\mathrm{Mpc}^{-1}$, represented in the rightmost column. From the panel in the first row it turns out that the local DM δ and the δ+T-web models fail to reproduce the reference bispectrum. The PT-web model improves the results of such models, but it still does not achieve the target level of accuracy. Finally, the I-web bias model is the one that meets the accuracy requirement to state that the bias is fully characterized at the investigated scales. The bispectrum of the mock catalog produced with I-web replicates the reference bispectrum with unprecedent accuracy. This becomes even clearer by looking at the $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$ ratios in the second row.

The third row indicates that the significance of the I-web bispectrum signal is within 1σ with respect to the reference bispectrum.

The bottom row, in the end, highlights two important and interesting facts:

- the significance of the δ+T-web model with respect to the DM δ one is $\sim 1.5\,\sigma$ in the wings (i.e. in the intervals $\theta/\pi < 0.25$, $\theta/\pi > 0.75$). This finding is consistent with the claim of Pellejero-Ibañez et al. (2020), which tested the δ+T-web model with the same outcome and regarded such result as a possible signature of non-local halo bias;
- the significance of the I-web model with respect to the DM δ local bias model is $\sim 4.8\,\sigma$ in the wings, which we regard as a clear detection of non-local bias.

As anticipated before, the PT-web model improves the DM δ local bias model and the δ+T-web model. However, the accuracy that such model attains in the bispectrum is not as high as the one achieved by I-web. Therefore, one could wonder which is in practice the difference between such models. Since PT-web consists in the set of generating functions $\{\delta,\, I_4,\, I_5\}$, it accounts for terms such as λ_i^2 and λ_i^3, but not for cross-terms such as $\lambda_i\lambda_j$, with $i \neq j$. I-web, instead, includes terms which mix up the eigenvalues through I_2 and I_3, and their full non-linear expansion. Up to our understanding, this difference is the reason why I-web models the anisotropic clustering more accurately than PT-web. In fact, as already stated previously I-web guarantees the full modelling of δ, s^2, the ellipticity e and the prolateness p and the partial modelling of s^3 and $\delta_L s^2$.

Finally, we investigate the importance of including explicitly the amount of information encoded in the 2nd and 3rd perturbative orders respectively. To this end, we build a proper numerical experiment. We run `BAM` with the following models:

- PT-web: $\{\delta,\, I_4\}$, where I_4 includes quadratic terms such as λ_i^2 (but not cubic terms);

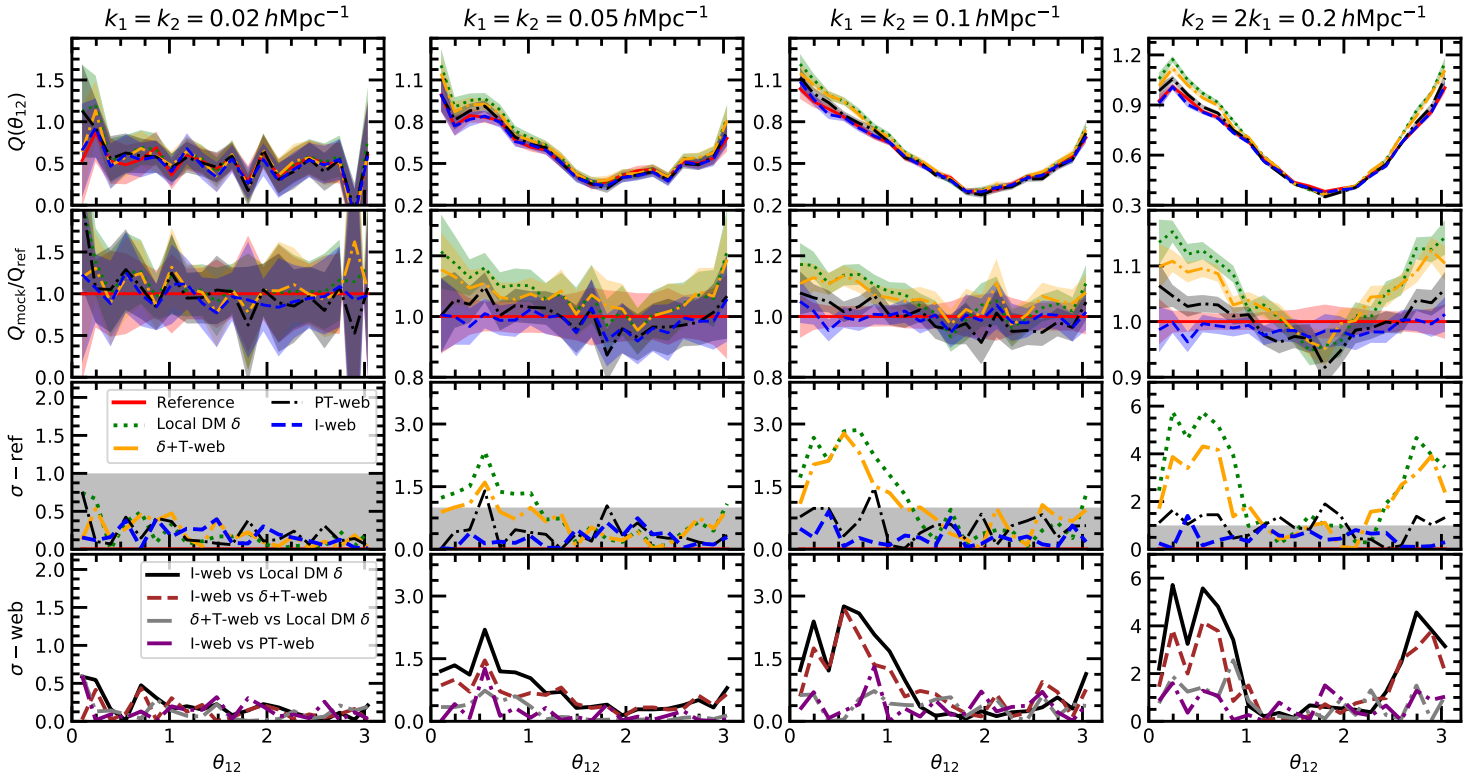

Figure 12: The first row shows reduced bispectra of the reference N-body halo catalog (red solid) and the BAM mocks with the bias relation including: only δ (green dotted), δ+T-web (solid orange), PT-web: $\{\delta, I_4, I_5\}$, and the I-web $\{\delta, I_2, I_3\}$. The second row shows instead, with the same colours, $Q_{\mathrm{mock}}/Q_{\mathrm{ref}}$. The third row shows the sigma difference: $\sigma_{\mathrm{ref}}(\text{X-web}) = |\langle Q_{\text{X-web}}\rangle - \langle Q_{ref}\rangle|/(\sqrt{\sigma^2_{\mathrm{ref}} + \sigma^2_{\mathrm{web}}+})$, with σ being the standard deviation extracted from the 300 MINERVA reference halo catalogs. The last row shows the difference between two X and Y BAM mocks $\sigma_{\mathrm{web}}(\mathrm{X},\mathrm{Y}) = |\sigma_{\mathrm{ref}}(\text{X-web}) - \sigma_{\mathrm{ref}}(\text{Y-web})|$, as indicated in the legend. The largest evidence for non-local bias is found for the I-web: $\sigma_{\mathrm{web}}(\text{I-web}, \delta)$.

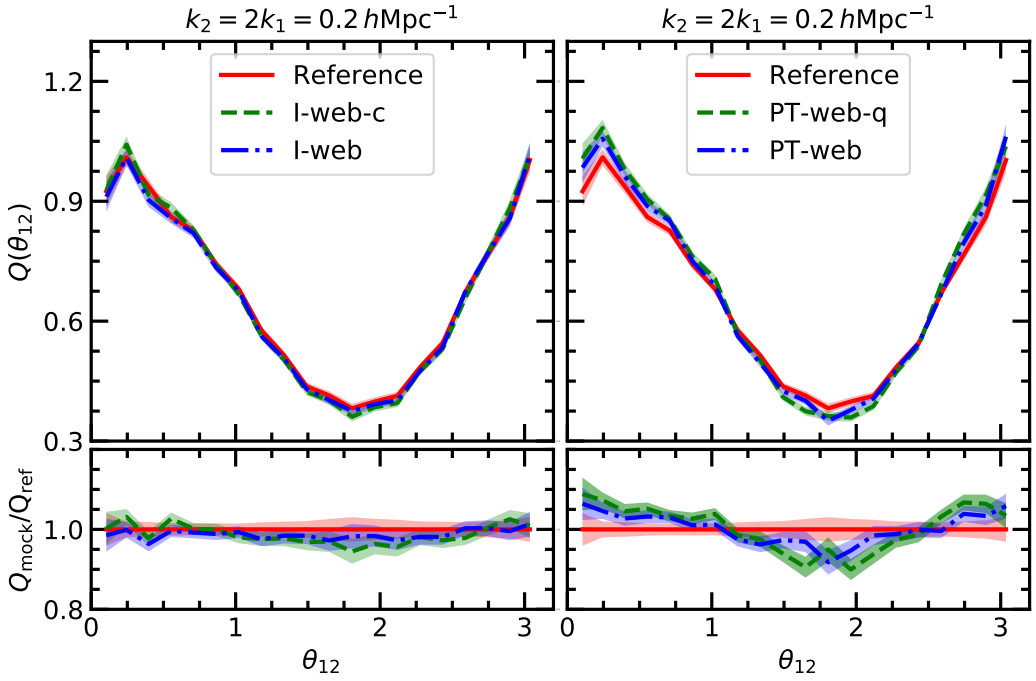

Figure 13: In the top row reduced bispectra and in the bottom row ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$ at configuration $k_1 = 0.1$, $k_2 = 0.2\,h\mathrm{Mpc}^{-1}$ in Fourier space. In the panels on the left, a comparison between reference bispectrum (red solid line) and mock bispectra, produced respectively with the I-web (blue dash-dotted line) and I-web-c (green dashed line). On the right, the same plots but with PT-web-q (green dashed line) and PT-web (blue dash-dotted line). The results tell us that I-web-c and I-web produces mock bispectra which are in good agreement. Instead, PT-web adds a significant amount of information on the spatial distribution of halos with respect to PT-web-q.

- PT-web-q: $\{\delta, I_4, I_5\}$, where I_4 and I_5 include respectively quadratic terms such as λ_i^2 and cubic terms such as λ_i^3.

Figure 13 presents the results of this final experiment. The right panels show the comparison between the mock bispectra obtained with PT-web (blue dash-dotted line) and PT-web-q (green dashed line), while the left panels show again the comparison between the mock bispectra obtained obtained with I-web (blue dhas-dotted line) and I-web-c (green dashed line). The reference bispectrum is plotted as a red solid line. The top row shows reduced bispectra $Q(\theta_{12}, k_1, k_2)$, while the bottom row shows ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$ between mock and reference bispectra. It turns out that the PT-web model reproduces more accurately the reference 3-point statistics than PT-web-q. This means that the 3rd order term I_5 still adds a significant amount of information to a model in which we restrict the dependencies to second order only.

The analysis presented so far makes it clear that I-web is the only bias model which produces mock catalogs whose bispectrum is fully compatible with the reference bispectrum at all the presented scales. However, it is necessary to make a further clarification.

In our calculations we have implicitly assumed that the mock catalogs produced by BAM with different bias models arise from different initial conditions. In other words, we have assumed that the cosmic variance arises from independent samples. This is not true, because the different halo mock catalogs we have produced are tracers of the same dark matter field, i.e. we are producing different halo mocks of the same realization. This makes the estimations of errors presented so far quite conservative. To refine the assessment of errors, we performed an additional statistical analysis, described in what follows.

In general, the computation of the variance of bispectrum in the full non-linear regime has not a well-established analytical solution. Moreover, in our analysis we have been using a non-linear mapping of tracers onto the dark matter field and the original MINERVA dark matter field has been downsampled from a $N = 1000^3$ to a $N = 500^3$ cells mesh. All these facts make the estimation of the bispectrum variance a complicated task.

In order to solve this problem we adopted the following idea. Let $Q(\theta_{12}, k_1, k_2)$ be the reduced bispectrum, given in the form of an ensemble of binned measurements. Let us also assume that all the values of Q are jointly Gaussian-distributed, i.e. they can be drawn from a multi-dimensional Gaussian probability distribution, whose dimensionality is given by the number of bins in the bispectrum. Under these assumptions, let us denote with μ_Q the vector representing the mean bispectrum we would estimate if we had available a set of measurements for Q. Finally, let us call $\mathcal{C}$ its covariance matrix. Notice that at this point μ_Q and $\mathcal{C}$ are in general unknown. Furthermore, let us look not at reduced bispectra, but at differences $Q_{\rm mock} - Q_{\rm ref}$ between mock and reference bispectra. This is useful in order to smooth the features of bispectra introduced by cosmic variance. Under these assumptions Gaussian processes (GPy, since 2012; Rasmussen & Williams, 2005) can be used to estimate both μ_Q and $\mathcal{C}$. The covariance matrix $\mathcal{C}$ is computed as:

$$\mathcal{C}_{ij} = \sigma_{\mathcal{C}}^2 \exp\left[-\frac{(\theta_{12,i} - \theta_{12,j})^2}{2\, l_{\mathcal{C}}^2}\right] + \sigma_\epsilon^2\, \delta_{ij}^{\rm K}\,, \tag{52}$$

where $\delta^{\rm K}$ is the Kronecker delta, l_c is a parameter which controls the smoothness of the function, σ_c its amplitude and σ_ϵ is the noise expected from the measurements. These parameters can be estimated by minimizing the likelihood:

$$\log\, P(\boldsymbol{Q}|\boldsymbol{\theta}_{12}, \mathcal{C}) = -\frac{1}{2}\boldsymbol{Q}^{\rm t}\,\mathcal{C}^{-1}\,\boldsymbol{Q} - \frac{1}{2}\log(\det\mathcal{C}) - \frac{N}{2}\log(2\pi)\,, \tag{53}$$

where N is the size of the array Q. In this way, new errors bars are computed. Figure 14 shows residuals $Q_{\rm mock} - Q_{\rm ref}$, with the newly computed error bars as red shaded areas and the error bar computed by averaging the bispectra of the 300 MINERVA realizations as a gray shaded area. The new error bars turn out to be smaller and this confirms that the first estimations were indeed a bit conservative. In this picture, the bispectrum obtained with the I-web model is the unique which is compatible with the reference bispectrum within error bars.

In order to assess the validity and the significance of each of the models at the various scales in this new perspective, we adopt the χ^2 statistics. Precisely, we adopt the following outline:

- first, we compute $\chi^2_{\rm dof} = \chi^2/{\rm dof}$, where dof indicates the degrees of freedom, which in this case are the number of bins adopted in θ_{12};
- we compute the probability $P(X > \chi^2_{\rm dof})$ for a model to produce a $\chi^2 = X$ larger than $\chi^2_{\rm dof}$;
- we reject the model if the probability $P(X > \chi^2_{\rm dof}) < 1$, because we assume that the measured χ^2 is not drawn from the same distribution, if its probability is not within the 99% of the distribution.

The results of this analysis are reported in Table 2. By looking at them we conclude that I-web is the unique bias model that cannot be rejected at any of the considered scales.

Then, we also estimate which is the *evidence* for preferring one bias model to another model. This is assessed by means of the Bayes factor of the $\chi^2/{\rm dof}$.

Given two models M_1, M_2 and a set of data D for which one has to perform a model selection problem, the Bayes' factor is defined as:

$$K = \frac{P(D|M_1)}{P(D|M_2)} = \frac{P(M_1|D)}{P(M_2|D)}\,\frac{P(M_1)}{P(M_2)} \quad , \tag{54}$$

where the first equality is a definition and the second equality holds due to the Bayes' theorem. Since the priors of the models $P(M_1) = P(M_2)$, because we have assumed flat priors in absence of extra knowledge, K reduces to the ratio $K = P(M_1|D)/P(M_2|D)$, i.e. to the ratio of the likelihoods computed in equation 53. In this perspective, we compute the Bayes factors needed to understand which is the evidence for preferring I-web with respect to all the other models at the various scales.

Then, Bayes factors are interpreted in terms of evidence of preferring the model M_1 with respect the model M_2 following Jeffreys (1939):

- $K < 1$: negative evidence (supports M_2);
- $1 < K < 3.2$: no evidence (N);

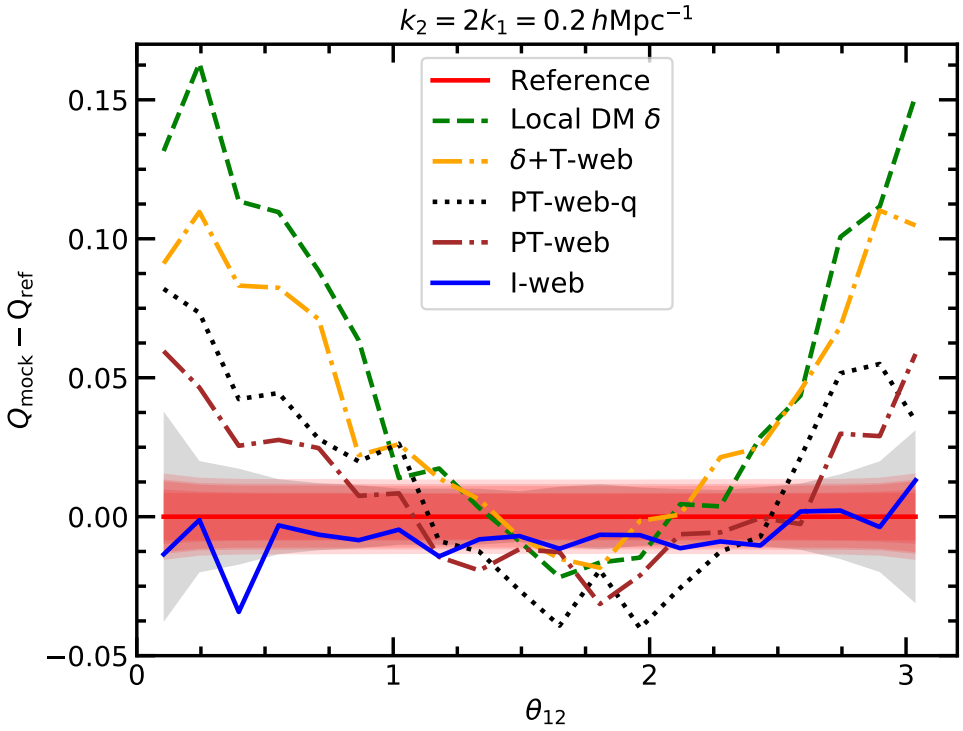

Figure 14: Difference $Q_{\mathrm{mock}} - Q_{\mathrm{ref}}$ between the reduced bispectrum of the mock catalogs and the reference halo catalog (red solid line). The mock catalogs has been obtained including in the bias respectively local DM δ (green dashed line), δ+T-web (yellow dash-dotted line), PT-web-q (black dotted line), PT-web (brown dash-dotted line) and I-web (blue dashed line). Red shaded areas stand for the newly computed error bars, while the gray shaded area represents the error computed by averaging the 300 realizations of the `MINERVA` halo catalogs. The bispectrum obtained with I-web is the unique which is compatible with the reference within the new error bars.

- $3.2 < K < 10$: substantial evidence (Y1);
- $10 < K < 32$: strong evidence (Y2);
- $32 < K > 100$: very strong evidence (Y3);
- $K > 100$; decisive evidence (Y4).

The results of these calculations are shown in Table 3. On large scales there is no evidence for preferring one model with respect to another one. This is consistent with the fact that on such scales noise dominates, error bars are large and there is no visual evidence that one of the mock bispectra fits better the reference bispectra with respect to the other mocks. However, when going to smaller scales there starts to be evidence of the preference of I-web to the other models. Indeed, at $k_1 = 0.1, k_2 = 0.2\,h\mathrm{Mpc}^{-1}$ there is very strong evidence (Y3) for preferring I-web to PT-web and PT-web-q and decisive evidence (Y4) for preferring I-web to local DM δ and δ+T-web. This fully confirms the preliminary conclusions we had drawn by looking at Figure 12.

In conclusion, we claim that at these large scales the 3-dimensional distribution and clustering of DM halos is fully determined by gravity, through the geometry of the gravitational potential and the anisotropic clustering of the cosmic web, the latter being defined by the I-web. We claim a clear detection of non-local bias by looking at the 3-point statistics in Fourier space and by noticing that the I-web model is able to reproduce the reduced bispectrum at $k_1 = 0.1, k_2 = 0.2\,h\mathrm{Mpc}^{-1}$ scales of the reference halo catalog with unprecedent accuracy. All these results highlight the fact that a proper modelling of the cosmic web is fundamental and that it is a key ingredient to determine the mapping of the DM halos field onto the underlying DM field.

In the following chapter we will extend the bias mapping idea to the baryonic Physics of the intergalactic medium and examine which is the role played in that case by the cosmic web, both in its dark matter and its baryonic components. This will be done in the perspective of assessing which are the crucial variables to be taken into account in the bias model in order to accurately reproduce the 3-dimensional distribution and clustering of the main properties of the intergalactic medium (ionized gas density, HI number density, temperature and optical depth).

Before concluding this chapter, it is worth mentioning some possible extensions of this work.

4 TOWARDS SMALLER SCALES

Looking at the inspiring bias expansion in equation 38, in this study we have neglected terms which can be built from the peculiar velocity fields

Bias model	Bispectrum configuration							
	$k_1 = k_2 = 0.02$		$k_1 = k_2 = 0.05$		$k_1 = k_2 = 0.1$		$k_1 = 0.1$ & $k_2 = 0.2$	
	$P(X > \chi^2)$	MR	$P(X > \chi^2)$	MR	$P(X > \chi^2)$	MR	$P(X > \chi^2)$	MR
Local DM δ	0.2	**Y**	~ 0	**Y**	~ 0	**Y**	~ 0	**Y**
δ+T-web	36.6	**N**	~ 0	**Y**	~ 0	**Y**	~ 0	**Y**
PT-web-q	9.0	**N**	~ 0	**Y**	0.04	**Y**	~ 0	**Y**
PT-web	26.1	**N**	1.2	**N**	~ 0	**Y**	~ 0	**Y**
I-web	38.3	**N**	19.3	**N**	45.8	**N**	1.2	**N**

Table 2: This table shows the probability $P(X > \chi^2_{\rm dof})$ for a model to produce a $\chi^2 = X$ larger than $\chi^2_{\rm dof}$ for each of the considered bias model and at each of the investigated scales. Each $P(X > \chi^2_{\rm dof})$ is provided with its MR (Model Rejection), where a model is rejected if it produces $P(X > \chi^2_{\rm dof}) < 1$, i.e. if its probability is not within the 99% of the distribution. The results clearly show that the unique model that cannot be rejected at all scales is the I-web one. See Kitaura et al. (2020) for a more detailed version of this Table.

Bias model	Bispectrum configuration							
	$k_1 = k_2 = 0.02$		$k_1 = k_2 = 0.05$		$k_1 = k_2 = 0.1$		$k_1 = 0.1$ & $k_2 = 0.2$	
	BF	E	BF	E	BF	E	BF	E
I-web vs Local DM δ	1.71	**N**	48.34	**Y3**	786.13	**Y4**	3823632	**Y4**
I-web vs δ+T-web	1.01	**N**	76.90	**Y3**	710.38	**Y4**	231053	**Y4**
I-web vs PT-web-q	1.21	**N**	6.04	**Y1**	2.02	**N**	61.84	**Y3**
I-web vs PT-web	1.06	**N**	1.34	**N**	2.95	**N**	38.97	**Y3**

Table 3: This table shows the Bayes factors (BF) for each of the considered bias model comparison and at each of the investigated configurations on Fourier space. For each X vs Y model comparison, the BF is provided with the evidence (E) of the preference of the X model with respect to Y. The evidence grades are drawn from Jeffreys (1939) (N: No evidence, Y1: Substantial, Y2: Strong, Y3: Very Strong, Y4: Decisive). The results show that on large scales there is no evidence for preferring one model to another one. However, at the smaller considered configuration, there is strong ($Y3$) evidence for preferring I-web to δ-only and δ+T-web models and decisive evidence ($Y4$) for preferring I-web to PT-web and PT-web-q models. See Kitaura et al. (2020) for a more detailed version of this Table.

(e.g. the invariants of the velocity shear tensor) and short-range non-local bias (e.g. $\nabla^2\delta$) terms. According to our analysis, we achieve a complete description of the halo bias on large scales, e.g. $k_1 = 0.1, k_2 = 0.2h\,\mathrm{Mpc}^{-1}$ in Fourier space. Therefore, we argue that the terms we neglected are probably not important at the scales we have analyzed.

However, the current halo bias picture predicts a stochastic scale-dependent bias. Hence, the fact that such neglected terms might become important at other scales is not out of the question. On the contrary, when going towards smaller scales (e.g. $k_1 = 0.3, k_2 = 0.6\,h\mathrm{Mpc}^{-1}$), I-web may fail to fully constrain the spatial distribution of DM halos. Concretely, we expect that the reduced bispectrum $Q(\theta_{12}, k_1, k_2)$ of the mock catalogs obtained with I-web will no longer fit the reference bispectrum as accurately as it does on large scales. Instead, other terms, such as velocity or short-range non-local terms, may become crucial to accurately replicate the 3-point statistics of the reference simulation in the mock catalogs.

We plan to investigate this aspect in future works.

5 BAM ON HYDRODYNAMICAL SIMULATIONS

This chapter presents how BAM is applied to a hydrodynamical simulation to learn the mapping relation between the intergalactic baryonic tracers (i.e. gas density, HI number density, temperature and optical depths for Ly-α photons) of the dark matter field and its cosmic web.

Hitherto, we have considered the bias between dark matter halos and the underlying dark matter density field. This is a well established and studied relationship in the literature. However, we need to extend the concept of bias to the functional relationship between baryonic quantities. This relation might be very complex, non-linear and non-local. Nonetheless, as long as there is a certain correlation between the considered quantities (for example between the ionised gas and the dark matter density), it is possible to define some bias expressed as a general tensor. When such a relation has been identified, it can be extracted and calibrated numerically with BAM.

We follow here the same approach adopted for DM halos. First, BAM extracts the stochastic bias relation from two fields of the reference simulation and calibrates it with the kernel. Second, we use BAM with the same base reference field used for the calibration (in principle DM, but it may be any of the other fields) and the calibrated bias relation to reproduce maps of the tracer field.

The reference simulation this time is no longer a DM-only cosmological simulation, but a hydrodynamical simulation, which accounts for baryonic physics and solves both gravity and hydrodynamics. In order to be able to do Cosmology, the volume of the simulation box must be large enough. As already pointed out in chapter 4, there are not many research groups in the world running such simulations. We make use of a simulation of volume $V = 100^3\, h^{-3}\mathrm{Mpc}^3$, run by the Theoretical Astrophysics research group of Osaka University, and led by Prof. Kentaro Nagamine. The details of the simulation are presented in section 1.

As anticipated, a bias relation between two fields is expected to exist when such quantities have a significant cross-correlation. At this point, one could wonder which among the considered quantities are sufficiently correlated to justify the existence of a bias and why.

We can answer the first question by looking at the cross-correlation between some of the fields. Figure 15 shows the normalized cross-spectra $C(k) = \hat{C}(k)/\sqrt{P_1(k)P_2(k)}$, where $\hat{C(k)}$ is the cross-spectrum of the two fields considered and $P_1(k)$ and $P_2(k)$ are respectively their power spectra. The plot tells us that:

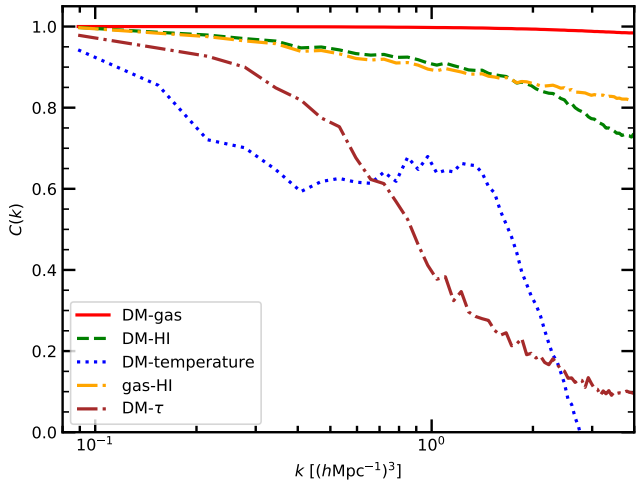

Figure 15: Normalized cross power spectra $C(k)$, as a function of wavenumber k, of a number of pairs of reference fields. The gas density is the field which correlates more with the DM density: $C(k) > 95\%$ at all scales. The HI number density correlates in a similar way with DM and gas density fields at large scales, but on small scales it correlates more with gas density than DM density (DM-HI: $C(k) \sim 75\%$, gas-HI: $C(k) \sim 85\%$). The DM-temperature and DM-τ cross-correlations are high on large scales, but abruptly drop on small scales.

- DM and gas density have a very high cross-correlation, $C(k) \sim 100\%$ at large scales and $C(k) > 95\%$ at small scales;
- HI has a lower correlation with DM density than the one the gas has, but still retains a considerable correlation, $C(k) > 70\%$, at all scales;
- the DM-HI and gas-HI cross-correlations are similar on large and intermediate scales, but on small scales the gas-HI $C(k)$ is higher than DM-HI $C(k)$ ($C(k) \sim 75\%$ vs $C(k) \sim 85\%$);
- the DM-temperature and DM-τ cross-correlations are high at the largest scales, but abruptly drop at small scales, becoming either vanishing (DM-temperature) or very low (DM-τ: $C(k) \sim 10\%$).

These calculations support the idea that a bias relationship between DM and respectively gas and HI density fields is well-justified. The temperature and the optical depth appear instead controversial, suggesting a well-defined large scale bias, but requiring a non-trivial treatment on small scales. This aspect will be discussed more in details in the dedicated sections.

The physical reason why such fields correlate can be understood with the following considerations:

- the models of structure formation predict that baryonic overdense regions are more likely to be found in correspondence of DM overdense regions;
- the temperature is determined (as will be explained in more details in section 2.3) by the photo-ionization equilibrium between ionization and recombination of intergalactic gas. In other words, the temperature is fully constrained by the relative abundance of ionized gas and HI densities and the UV photon background. We already noticed in Figure 15 that the temperature field does not correlate much with the DM field on small scales. However, on large scales it does and is expected to track the distribution of DM in the overdense regions;
- the optical depth for Ly-α photons is computed by properly integrating the HI number density along the line-of-sight (see section 2.4 for the details).

These arguments are qualitatively confirmed by the visual comparison of slices of the reference fields presented in Figure 16. By looking at that, one can easily track the same patterns defined by the overdense regions, common to all the fields.

Under these circumstances, we assume that the IGM fields can all be considered biased tracers of the underlying DM field and that eventually the concept of bias relation can be extended to pairs of baryonic quantities. In other words, we assume that a proper mapping between the baryonic fields and the DM field, or just between baryonic fields, exists and can be extracted from the reference simulation. When the correlation between two fields is not high enough to define a functional dependence which constrains their spatial distributions, the method does not converge properly to the reference power spectrum. In such cases, we will see that the code cannot converge properly or, in extreme cases, does not converge at all.

In order to account for the thermodynamical relationships between the baryonic fields, we employ the following mapping procedure (let us call *X field* the base field to which the kernel is applied and *Y field* the tracer field of which we want to produce the mocks):

- the gas density is mapped onto the dark matter field, i.e. X=DM density and Y=ionized gas density;
- the HI is mapped onto the gas density field, i.e. X=gas density and Y=HI number density;
- the temperature is mapped onto the dark matter field i.e. X=DM density and Y=temperature. In this basic case the code suffers convergence problems. For this reason, we include in the bias description also gas and HI densities at the same time, in order to add information on the relative abundance of the two. As anticipated above, this

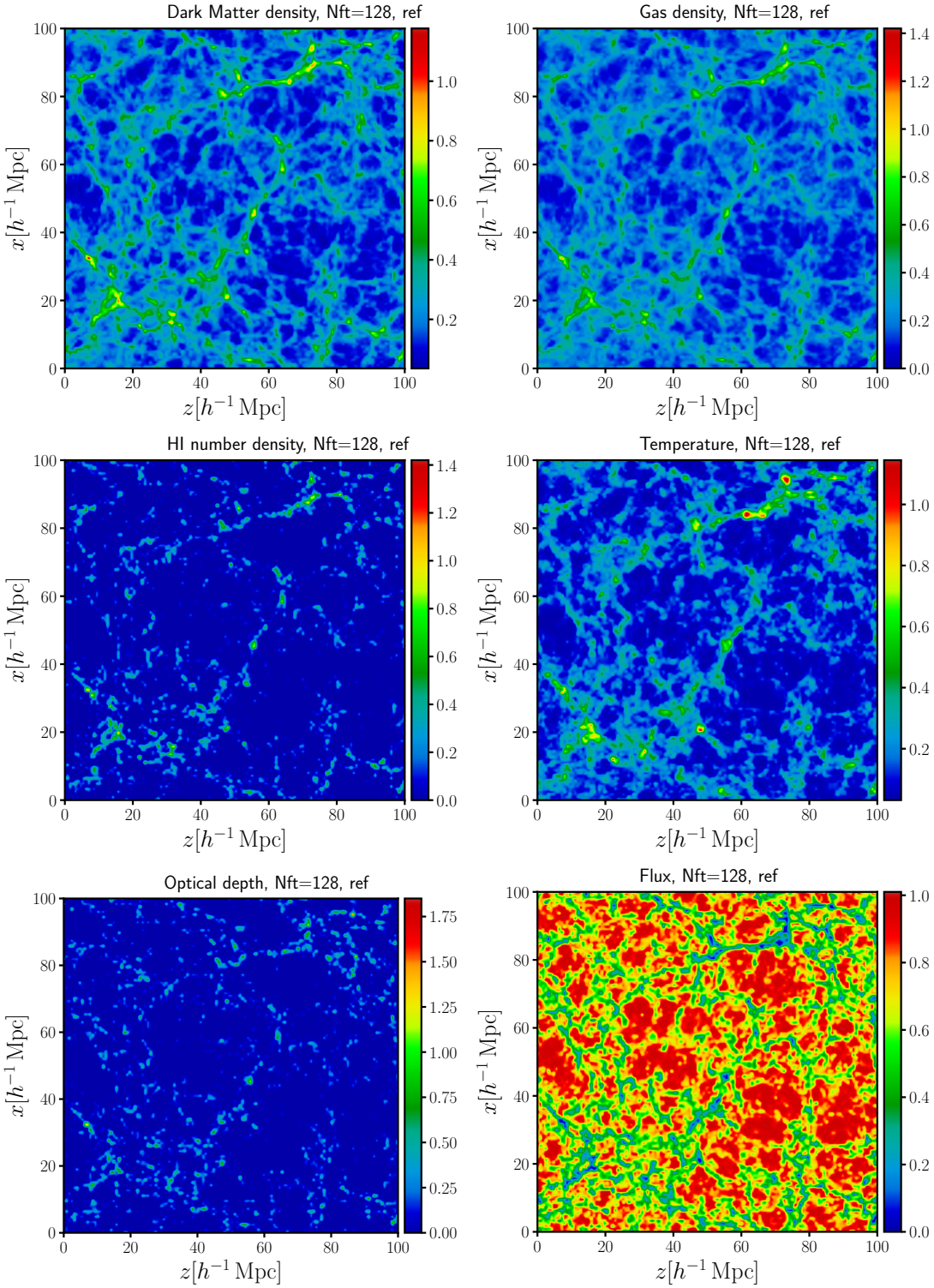

Figure 16: Slices of the reference simulation fields, extracted at the same position of the mesh. In the upper left panel, a slice of the DM density; in the upper right panel, a slice of the ionized gas density; in the mid left panel a slice of the HI number density; in the mid right panel a slice of the temperature; in the lower left panel a slice of the optical depth; in the lower right panel a slice of the Ly-α fluxes.

helps to constrain the temperature, according to the photo-ionization equilibrium condition;

- the optical depth is at first mapped onto the DM field, leading to an unsuccessful reproduction of the reference simulation. Then, some modifications applied to the dark matter field and described in details in section 2.4 lead to an improvement of the mapping. The final solution is still not efficient enough to provide a mock power spectrum which fulfills the convergence requirement. We leave this aspect as an open problem;

The calibration procedure is described in detail in section 2.

We are here in a situation which is different from the case of dark matter halos. In that case, indeed, we only have the DM density and DM halo fields available, therefore no other mapping other than the one of halos onto the DM field is possible. In the case of a hydrodynamical simulation, we are in a more privileged position. Indeed, we have more than one tracer field and we can also exploit the mapping between baryonic fields.

While in the calibration we have all the reference fields available, in the practical application of doing mocks we will not. Instead, we can rely only on the DM field produced with ALPT and the calibrated bias probability distributions. To map e.g. the HI number density, we will not have the reference gas density field, but a gas density mock catalog instead. And still, when we will sample temperature mocks, we will need to make use of the mock gas and HI density fields in the bias.

Therefore, on one hand adopting the mapping procedure described above potentially allows to gain information with respect to mapping all the tracers onto the DM field.

On the other hand, if such mapping really improves the results, it is crucial to calibrate the bias in such a way to reproduce very accurately the 2- and 3-point statistics of the reference simulation. Otherwise, a mock catalog may lose accuracy and precision when another mock catalog is used to map it, instead of a reference field. We do not analyze this aspect here, but one must keep in mind that this mapping procedure needs to be tested also in a realistic context of mock production, as an extension of this thesis.

1 THE REFERENCE SIMULATION

In this section I introduce the details of the reference simulation used to learn and calibrate the bias mapping relation.

The simulation has been run with the cosmological smoothed-particle hydrodynamics (SPH) code `Gadget3-OSAKA` (Aoyama et al., 2018; Shimizu et al., 2019), a modified version of the popular N-body code `Gadget2` (Springel, 2005). It embeds a volume $V = 100^3\,h^{-3}\mathrm{Mpc}^3$, large enough to perform

cosmological studies, and a number of particles $N = 2 \times 512^3$ of mass $M = 5.36 \times 10^8 h^{-1}\mathrm{M}_\odot$ for the DM particles and $M = 1.00 \times 10^8 h^{-1}\mathrm{M}_\odot$ for the gas particles. The softening length, adopted to prevent numerical divergences in gravity computations, is $\epsilon = 6h^{-1}$kpc. The number of neighbour particles for each SPH particle is $N_n = 128 \pm 8$ and the quintic spline kernel (Morris, 1996) is employed. The code contains also important refinements, like e.g. the density-independent formulation of SPH (Hopkins et al., 2013; Saitoh & Makino, 2013) and the time-step limiter (Saitoh, 2016).

The main baryonic processes which influence the Chemistry are photo-heating and photo-ionization under the UV background, described by the model in Haardt & Madau (2012), and radiative cooling. All these processes are accounted for by the `Grackle` library (Smith et al., 2017), which solves the Chemistry for atomic (H,D and He) and molecular (H_2 and HD) species.

The adopted cosmological parameters are the ones of Planck Collaboration et al. (2016), i.e.:

- baryon density parameter $\Omega_b = 0.049$;
- total matter density parameter $\Omega_m = 0.32$;
- cosmological constant density parameter $\Omega_\Lambda = 0.68$;
- Hubble constant $H_0 = 67\mathrm{km\ s^{-1}Mpc^{-1}}$;
- power spectrum index of scalar perturbations: $n_s = 0.9645$;
- normalization of density fluctuations: $\sigma_8 = 0.831$.

A number of additional important baryonic processes are taken into account:

- Star Formation: the gas particles which satisfy the conditions required for star formation produce star particles. Their mass is stochastically assigned in such a way to reproduce to Chabrier Initial Mass Function (Chabrier, 2003), in the mass range $0.1\mathrm{M}_\odot < \mathrm{M} < 120\mathrm{M}_\odot$;
- heavy metals, dust production and SN feedback, with the `CELib` library (Saitoh, 2016, 2017);
- stellar feedback (i.e. energy input from massive stars);
- dust evolution, following the model presented in Hou et al. (2017). In order to distinguish between dust belonging to galaxies and dust of the IGM, the code `P-STAR Groupfinder` (Springel et al., 2001) is used.

See Aoyama et al. (2018) and Shimizu et al. (2019) for a detailed description and formalism of all the mentioned baryonic processes.

The initial conditions are generated at redshift $z \sim 99$ with the `MUSIC` code (Hahn & Abel, 2011). The outputs of the simulations we use here are snapshots extracted at redshift $z = 2$. The provided intergalactic medium fields are:

- dark matter density field, in units of g/cm^3;
- ionized gas density, in units of g/cm^3
- neutral hydrogen (HI) number density, in units of cm^{-3};
- gas temperature, in units of K;
- optical depths τ for the Lyman-α photons (see chapter 2), computed along the 3 lines-of-sight given by the 3 coordinate directions x, y and z (Shimizu et al., 2019).

A special mention is to be dedicated to the computation of the optical depths τ. Given the numerical complexity and the time required for the integration along the lines-of-sight of all the directions of the simulation box (see section 2.4), the approach which is commonly adopted is to assume the Fluctuating Gunn-Peterson Approximation (see chapter 2).

However, thanks to the work performed by the Osaka research group, we can here rely on optical depth fields computed exactly (i.e. by integrating along the line-of-sight, see section 2.4). This not only allows to calculate very accurately the Lyman-α fluxes, but also to compare the FGPA approximation with the exact values of τ.

The results of the simulation are provided in a set of 3 different resolutions, i.e. in a set of 3 different number of cells per side adopted in the interpolation on the mesh, $N_{\text{cells}} = 128^3$, $N_{\text{cells}} = 256^3$ and $N_{\text{cells}} = 512^3$ respectively.

Notice that the case with the lowest resolution, i.e. $N_{\text{cells}} = 128^3$, is already a zoom-in high-resolution simulation, whose cell size is $l \sim 0.78\,h^{-1}$Mpc. All the cases allow to perform studies down to sub-Mpc scales, despite embedding a total volume which is enough for cosmological studies. We consider here only the case with the lowest-resolution. The results of the analysis will be presented in a paper we will submit in the near future.

Even though we have performed some tests also on the two cases with higher resolution, in order to prove the validity of the method, the detailed analysis of such cases goes beyond the scope of this work. They are to be investigated in further works, if needed.

2 MAPS REPRODUCTION WITH REFERENCE FIELDS

In this section we discuss the reproduction of the maps of the IGM properties. A number of specific bias models will be analyzed for each of the IGM properties. The method followed is the same of chapter 4, i.e. calibrating the bias relation, obtaining the converged power spectrum, compute the reduced bispectrum at different configurations in Fourier space and analyse

the results. Unlike the case presented in chapter 4, the resolution adopted here allows to go down to very small scales. Notice indeed that the Nyquist frequency is $k_{\rm nyq} \sim 4.02\,h{\rm Mpc}^{-1}$. On the other hand, because the volume is smaller, in large-scale configurations in Fourier space, like e.g. $k_1 = 0.1, k_2 = 0.2\,h{\rm Mpc}^{-1}$, the number of triangles making up the searched triangular configurations is small and this results in very noisy bispectra at such scales. For this reason, I perform bispectra studies on configurations of the type $k_1 = k_2 = \tilde{k}\,h{\rm Mpc}^{-1}$, where $0.5 < \tilde{k} < 1.0\,h{\rm Mpc}^{-1}$.
Like in chapter 4, for each of the studied properties of the IGM I will present first the visual comparison between the reference and the mock reproduction, then I will show the results regarding the convergence of the power spectra and subsequently I will present the results achieved in the reduced bispectrum. In doing this, I will take care of highlighting which are the statistical bias models which result in the best accuracy of the reproduced maps.
In the visual comparisons, the slices are extracted always at the same position of the simulation box and 5 contiguous slices are averaged together. Exact, for the transmitted Lyman-α fluxes, which are represented in linear scale, the field represented in the slice is first converted to overdensity δ and then represented in logarithmic scale $\log_{10}(2+\delta)$, where the latter transformation is employed in order to enhance the visual aspect of the overdense regions.

2.1 Ionized gas density

The first field we try to reproduce is the ionized gas density field. Following the procedure described above, we approach this study by mapping the gas density onto the dark matter density. By looking again at Figure 16, one can easily see that the spatial distribution of dark matter looks very similar to the one of ionized gas. Moreover, Figure 15 underlines that DM and gas density fields have a very high cross-correlation ($C(k) > 0.95$ at all the considered scales). Therefore, we choose to start with the gas density because it is the one which has the highest cross-correlation with the dark matter field.

We test the following bias models:

- local dark matter overdensity: $\{\delta\}$;
- short-range non-local bias: $\{\delta,\ \nabla^2\delta\}$
- long-range non-local T-web: $\{\delta,\ {\rm sgn}(\lambda_i)\}$;
- long-range non-local I-web: $\{\delta,\ I_2,\ I_3\}$;

λ_i are $\{I_2, I_3\}$ are respectively the eigenvalues of the tidal field tensor $\mathcal{T}$.

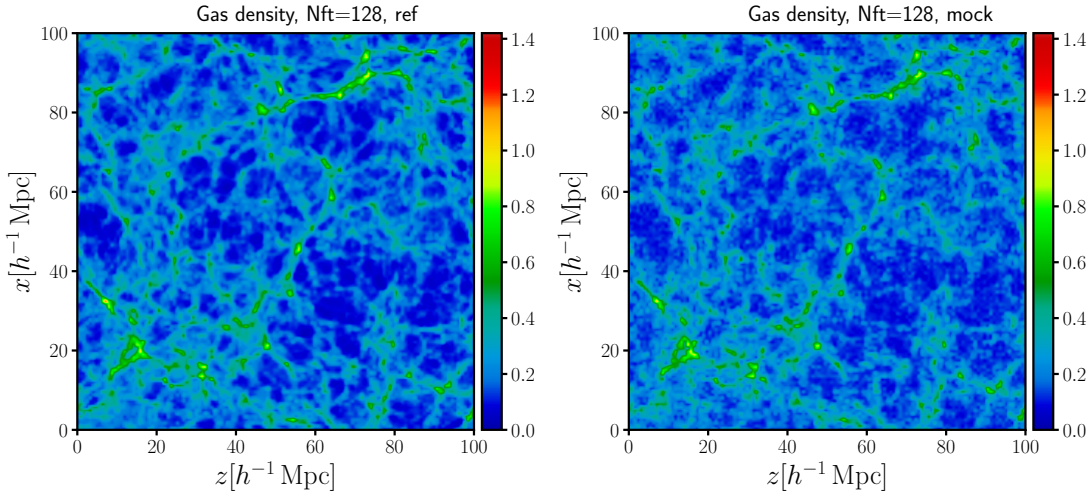

Figure 17: Slices of the gas density field, extracted at the same position of the mesh. On the left, a slice of the reference simulation box. On the right, a slice of the mock produced with BAM. The mock is obtained by calibrating the bias relation on the dark matter density field and by including the $\nabla^2\delta$ dependence in the bias model.

Contrary to our initial expectation, the I-web and T-web models turn out to be not as effective as in other cases, like halos or HI (see section 2.2 for the latter). We interpret this as the fact that in the mapping we are considering, the cosmic web dependence is not as crucial as in the cases we have just mentioned. Section 2.2 will make it clear that e.g. the I-web dependence is particularly important when we are dealing with halos or objects which are localized inside them. Instead, the analysis of 3-point statistics will reveal that the short-range non-local term $\nabla^2\delta$, which has not been tested in the halo bias, is important and allows to achieve high accuracy.

Qualitative assessment

Figure 17 shows a visual comparison between slices of the reference simulation box and the mock obtained with $\nabla^2\delta$. Beside some fragmentation and granulation of the smallest structures in the underdense regions, in general a good visual agreement is found. The overdense structures on large scales are in fact very well reproduced.

Quantitative assessment

Summary statistics: power spectrum In this section we analyze the results concerning the power spectrum of the gas density field, i.e. 2-point statistics in Fourier space.

Figure 18 shows the comparison between the reference (red solid line) and the gas mock power spectra obtained including in the stochastic bias respectively the local dark matter overdensity (green dashed line), the T-

Mock	$R(\%)$	$\sigma(\%)$
Local DM δ	0.56	0.59
T-web	0.78	0.84
$\nabla^2\delta$	0.59	0.71
I-web	0.62	0.75

Table 4: Percentage residuals R and standard deviation σ of ratios $P_{\rm mock}(k)/P_{\rm mock}(k)$ for the gas mocks obtained with the considered bias models.

web (yellow dash-dotted line), the short-range non-local term $\nabla^2\delta$ (blue dash-dotted line) and I-web (brown dotted line). The upper panel shows the power spectra $P(k)$, while the lower panel shows the ratio between mock and reference power spectrum $P_{\rm mock}(k)/P_{\rm ref}(k)$ for each of the considered bias model. Gray shaded areas stand respectively, from the darker to the lighter, for $1\%, 2\%, 5\%$ residuals. From a visual comparison it turns out that all the power spectra are well-converged and that do not deviate more than 2% at all scales.

A more quantitative analysis of the result is done by computing average residuals R and standard deviation σ of ratios $P_{\rm mock}(k)/P_{\rm mock}(k)$, up to $k \sim 0.7\,k_{\rm nyq}$, being $k_{\rm nyq}$ the Nyquist frequency. We limit the measure of R and σ to such range of frequencies because are the ones we can trust in a power spectrum measure, as are not affected by aliasing (Jing, 2005). The results are presented in Table 4.

At this point, we can conclude that fiducial convergence and high precision are ensured by all the bias models, but cannot drawn any other significant conclusion. Therefore, we need to move to 3-point statistics.

Summary statistics: bispectrum Figure 19 shows reduced bispectra of the gas mocks, compared to the reference bispectrum, for different configurations in Fourier space and for different tested bias models. The first and second row show respectively reduced bispectra $Q(\theta_{12})$ and ratios $Q_{\rm mock}(\theta_{12})/Q_{\rm ref}(\theta_{12})$, where shaded areas represent error bars. The third and fourth rows represent $\sigma_{\rm ref}$ and $\sigma_{\rm web}$ statistical significance, i.e. respectively the statistical significance of bispectrum obtained with a X cosmic web bias model with respect to the reference and the statistical significance of bispectrum obtained with a X cosmic web bias model with respect to the one obtained with a Y cosmic web bias model. The first is given by 50, while the second is instead computed as

$$\sigma_{web}(\mathrm{X},\mathrm{Y}) = \sigma_{\rm ref}(\text{X-web}) - \sigma_{\rm ref}(\text{Y-web}) \quad , \tag{55}$$

where differently from equation 51 we do not take here the modulus. Indeed, in equation 51 we computed the statistical significance $\sigma_{web}(\mathrm{X},\mathrm{Y})$ with the

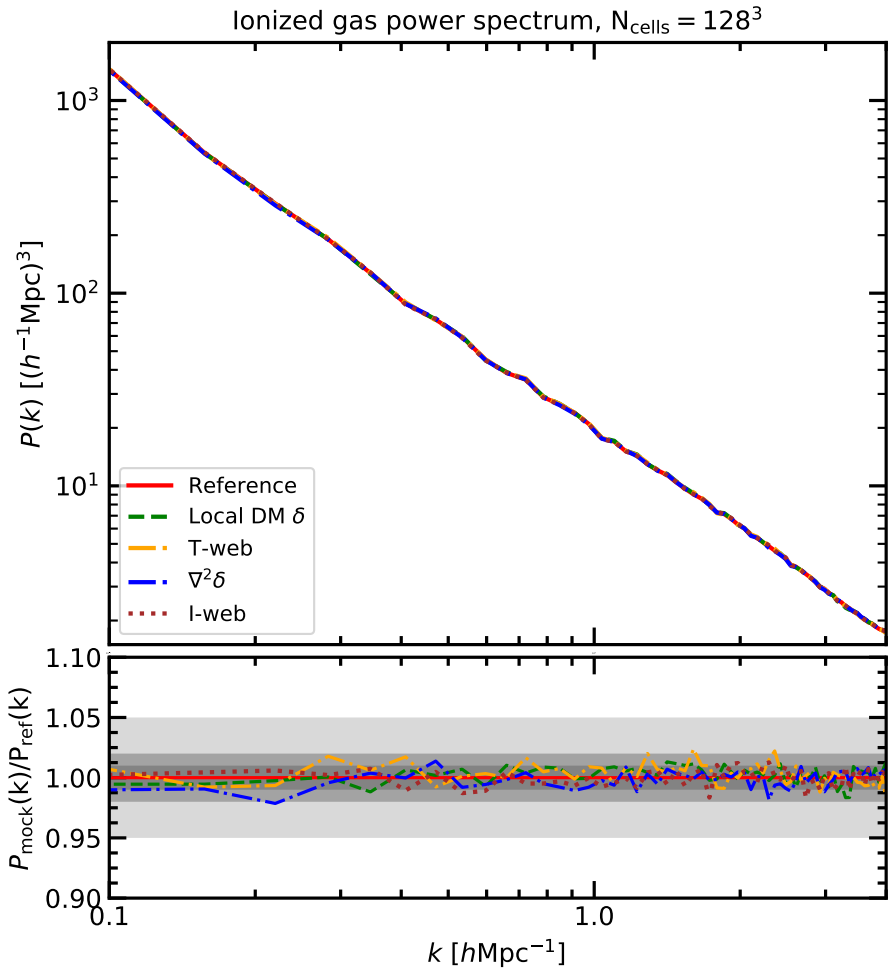

Figure 18: Comparison between reference (red solid line) and gas mock power spectra, the latter obtained including in the bias the dependence on the local dark matter overdensity (green dashed line), the T-web (yellow dash-dotted line), the short-range non-local term $\nabla^2\delta$ (blue dash-dotted line) and I-web (brown dotted line). The upper panel shows the power spectra $P(k)$, while the lower panel shows the ratio between mock and reference power spectrum $P_{\rm mock}(k)/P_{\rm ref}(k)$ for each of the considered bias model. Gray shaded areas stand respectively, from the darker to the lighter, for $1\%, 2\%, 5\%$ residuals.

awareness that Y was performing better than X in all the domain of θ. Here instead, given the large complexity of the problem, it is safer to compute $\sigma_{web}(X, Y)$ with its sign. We are interested in showing that a certain model yields a net gain of information with respect to another model. Thus, when $\sigma_{web}(X, Y)$ goes negative, it means that the preferential model is leading to a loss of information with respect to the other model. This may mildly happen e.g. in some narrow θ-intervals at the left border of the domain. However, if in the remaining part of the domain we detect a strong positive significance, we regard we can neglect the θ-region responsible for the loss of information.

Provided that we have only one realization available, errors on bispectrum measurements cannot be computed with the same procedure of chapter 4, where we computed error bars by averaging bispectra of the halo catalogs extracted from 300 different realizations. Instead, we adopt the following expression for the variance of (non-reduced) bispectrum σ_B^2 (Angulo et al., 2015; Fry et al., 1993; Scoccimarro, 2000; Scoccimarro et al., 1998), valid in Gaussian limit:

$$\sigma_B^2 \approx s_B \frac{k_f}{V_B} P(k_1)P(k_2)P(k_3) \quad , \tag{56}$$

where $s_B = 6, 2, 1$ for equilateral, isosceles and scalene configurations in Fourier space, $V_B \approx 8\pi^2 k_1 k_2 k_3 \Delta k$ and the fundamental mode is $k_f = 2\pi/\mathrm{Lbox}$, where Lbox is the physical size of the simulation box given in Mpc/h. The error on the reduced bispectrum $Q(\theta_{12})$ is then computed by error propagation. See Appendix B for a full-page version of Figure 19.

The results reveal that the most accurate reproduction of the gas density reference field is obtained by including in the stochastic bias the dependence on the short-range non-local bias term $\nabla^2\delta$. Instead, as anticipated above, accounting for the dependence on the cosmic web at long-range through T-web and I-web does not significantly improve the result, with respect to the basic local δ bias model. This may seem a bit counter-intuitive at this point, but will be clarified at the end of the analysis regarding HI.

Including $\nabla^2\delta$ in the bias is a novel aspect, because hitherto we have never had the necessity to account for such terms. Even though a better accuracy can still be achieved, we regard that $\nabla^2\delta$ allows to reproduce well the reference 3-point statistics, given the high complexity of the problem. The mock bispectrum obtained with such model is compatible within error bars, except for the region around $\theta \sim 2$. The statistical significance with respect to the reference is within $1\,\sigma$. Moreover, the fourth row indicates that a non-negligible statistical significance ($2 - 3\,\sigma$ at all scales) of $\nabla^2\delta$ with respect to local δ model (black dashed line) is detected. Following the lines of chapter 4, this is interpreted as an evidence of short-range non-local bias.

$\nabla^2\delta$ leads to a gain of information with respect to T-web (gray dash-dotted line) and I-web (purple dotted line) too.

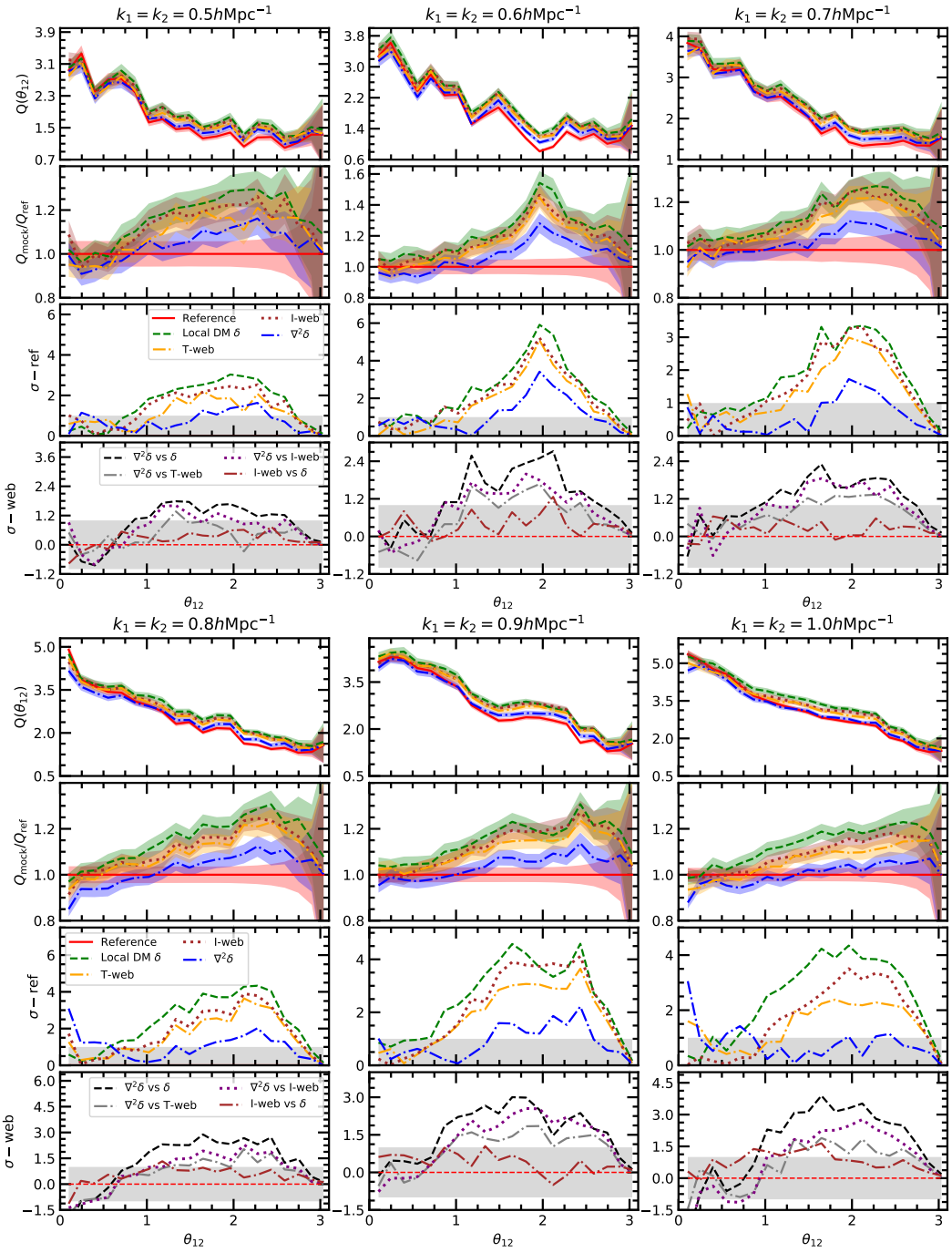

Figure 19: Reduced bispectra of reference field (red solid line) and gas mock catalogs, including in the bias respectively the dependence on local DM overdensity (green dashed line), the T-web (yellow dash-dotted line), the short-range non-local term $\nabla^2\delta$ (blue dash-dotted line) and I-web (brown dotted line). In the first row, reduced bispectra $Q(\theta_{12})$ and in the second row, ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$. Shaded areas represent error bars. In the third row and and fourth rows respectively the statistical significance $\sigma - \mathrm{ref}$ with respect to the reference and $\sigma - \mathrm{web}$ of a X bias model with respect to a Y one. Gray shaded areas stand for 1σ significance. The most accurate reproduction of the reference bispectrum is obtained including $\nabla^2\delta$ in the bias relation.

Finally, the brown dash-dotted line quantifies the eventual gain of information we get by assuming the dependence on I-web in the bias, with respect to local DM δ model. It turns out that such significance is within 1σ at all scales. We regard the amount of non-local information carried by I-web has not a significant impact at the scales we have considered.

Therefore we conclude that the spatial distribution of gas density is efficiently constrained by the geometrical configuration of the surrounding distribution of dark matter on short-range, while is not affected by its long-range cosmic web. Possible explanations for such fact is that either we are not looking at scales which are large enough to see the effect of the information carried by I-web; or the gas density, which unlike HI is massively present in the diffuse IGM outside collapsed objects, is not as sensitive as other tracers (e.g. halos and HI) to the long-range geometrical configuration of the dark matter cosmic web. Moreover, additional dependencies not investigated here, like the shear of the velocity field, may have an impact in the description of the 3-dimensional distribution and clustering of ionized gas.

We are here looking at relatively small scale configurations in Fourier space, where non-linearities are important. The bispectrum on larger scales becomes too noisy to establish the preference of one bias model with respect to other models, given limited the volume of our reference simulation. However, in a PT perspective, one could go on large scales and formalize a perturbative expansion in the spirit of McDonald & Roy (2009), expressing the ionized gas density as a function of local and non-local terms built from the dark matter density field.

2.2 HI number density

Following the procedure mentioned in the introductory part of this chapter, the final mapping of the HI number density is performed onto the ionized gas density. However, in order to prove that such mapping is more effective than a mapping onto the dark matter field, we consider here both the possibilities. We test the following bias models:

- local DM density: $\{\delta_{\rm DM}\}$;
- local gas density: $\{\delta_{\rm gas}\}$;
- DM I-web: $\{\delta_{\rm DM}, I_{2,\rm DM}, I_{3,\rm DM}\}$;
- gas I-web: $\{\delta_{\rm gas}, I_{2,\rm gas}, I_{3,\rm gas}\}$.

$\{I_{2,i}, I_{3,i}\}$ are the invariants of the tidal field tensor $\mathcal{T}_i$, built from the field i.

As pointed out in the introductory part of this chapter, in a realistic context of mock production one will not have the reference gas density field, but a gas mock catalog. Hence, if the mapping of HI on the reference

gas density turns out to better than the mapping on the reference DM, one should check that the same results are achieved also with respectively the mock gas density field and the ALPT DM density field. However, we assume here this is true (but will be tested explicitly in future works) and test the ideal mapping onto the reference fields.

According to Figure 15, the HI number density field highly correlates both with DM and gas density fields on large scales. However, on small scales the gas-HI cross-correlation is higher than the DM-HI cross-correlation by $5 - 10\%$. This fact supports the expectation that mapping HI on the gas field rather than on the DM field better constrains its spatial distribution. Mapping the HI onto the gas may benefit from accounting for the small scales baryonic processes which involve HI and gas but not dark matter.

The results presented in the following sections regarding 2- and 3-point statistics confirm these expectations. The mapping on the DM field works reasonably well, but the mapping on gas allows to achieve better accuracy. In particular, the I-web computed from the gas field constrains very accurately the 3-point statistics of HI. This is important mainly for two reasons:

- mapping the HI onto the gas density field allows to consistently account for the thermodynamical relations between such baryonic species. If, on the other hand, the DM field is used to map the HI, the consistency of thermodynamical relations between gas and HI mock mocks is not guaranteed;
- the best result is obtained including the gas I-web as non-local dependence in the bias. This is a novel aspect, because the I-web is extended for the first time to one of the baryonic tracer field of the DM distribution.

This is why in the end we choose to map preferentially HI on gas and not on DM. Moreover, once again the I-web model turns out to be a fundamental ingredient to describe the amount of information which the cosmic web embeds.

Qualitative assessment

The visual comparison presented in Figure 20 between slices of the reference HI field and the HI mock obtained with I-web$_{\text{gas}}$ shows that the mock is visually nearly indistiguishable from the reference. Even though the quantitative analysis will be presented in the following sections, such visual agreement is already a good qualitative indicator that we are properly reproducing the reference hydrodynamical simulation.

Quantitative assessment

Summary statistics: power spectrum In this section we analyze the results concerning the power spectrum of HI.

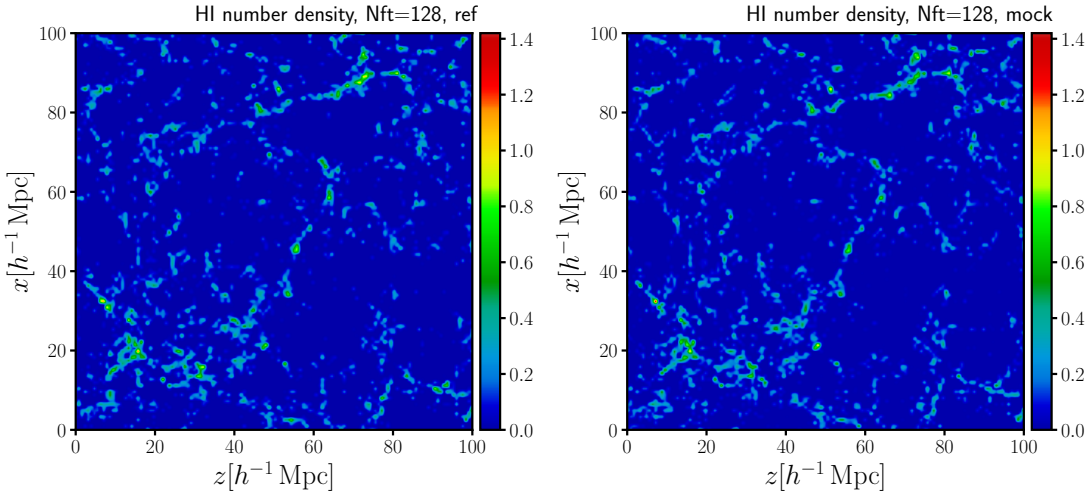

Figure 20: Slices of the HI number density field, extracted at the same position of the mesh. On the left, a slice of the reference simulation. On the right, a slice of the mock produced with BAM. The mock is obtained by calibrating the bias relation on the gas density field and by including the I-web model in the bias description.

Figure 21 shows the comparison between the reference (red solid line) and the mock power spectra obtained including in the stochastic bias respectively the local dark matter overdensity (green dashed line), the local gas overdensity (brown dotted line), the dark matter I-web (yellow dash-dotted line) and the gas I-web (blue dash-dotted line). The upper panel shows the power spectra $P(k)$, while the lower panel shows the ratio between mock and reference power spectrum $P_{\mathrm{mock}}(k)/P_{\mathrm{ref}}(k)$ for each of the considered bias model. Gray shaded areas stand respectively, from the darker to the lighter, for $1\%, 2\%, 5\%$ residuals.

From a visual comparison it turns out that all the mock power spectra are well-converged to the reference power spectrum. The lower panel shows that, except for the dark matter I-web case, the mock power spectra do not deviate more than 2% from the reference power spectrum at all scales. The mock power spectrum obtained with DM I-web loses accuracy at intermediate scales, where it tends to systematically exceed the reference.

A more quantitative analysis of the result is done by computing average residuals R and standard deviation σ of ratios $P_{\mathrm{mock}}(k)/P_{\mathrm{mock}}(k)$, up to $k \sim 0.7\,k_{\mathrm{nyq}}$, being k_{nyq} the Nyquist frequency. We limit again the measure of R and σ to such range of frequencies, in order to avoid aliasing. The results are presented in Table 5. The most precise results are obtained with the gas I-web.

Therefore, from a $P(k)$ analysis, the best mapping of HI turns out to be the one onto the gas density field which accounts also for the gas I-web in the bias. This fact is investigated more in details in the following paragraph, regarding the 3-point statistics.

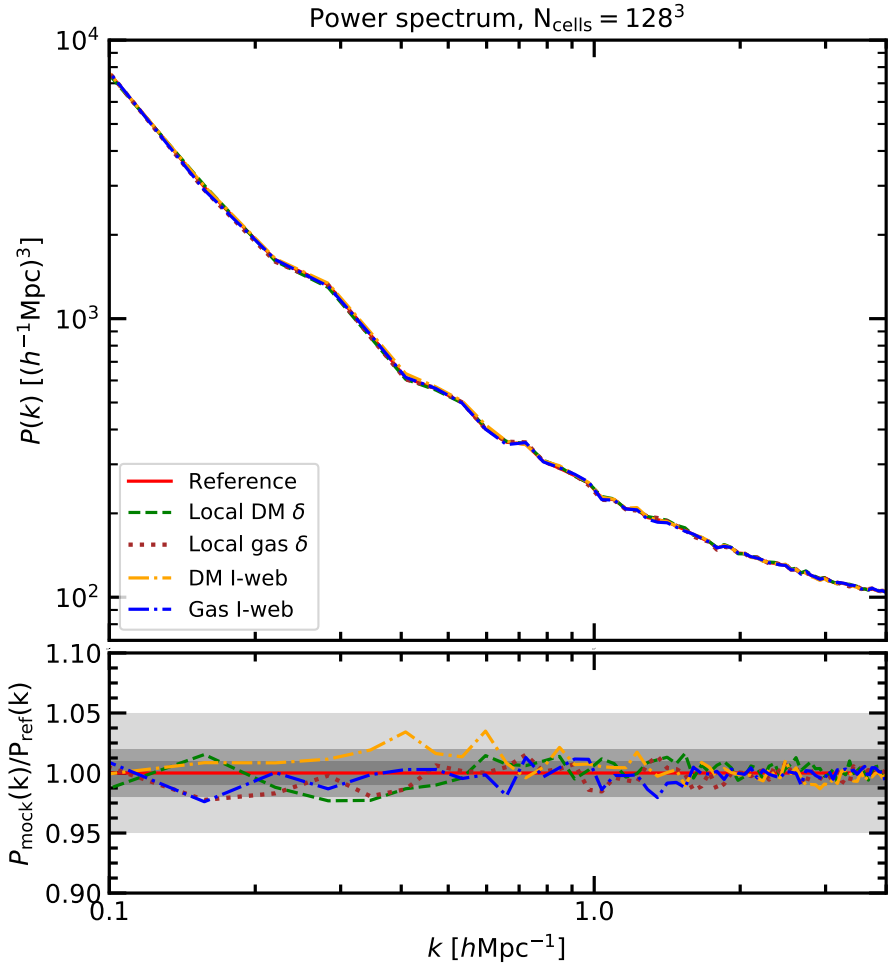

Figure 21: Comparison between reference (red solid line) and HI mock power spectra, the latter obtained including in the bias the dependence on the local dark matter overdensity (green dashed line), the local gas overdensity (brown dotted line), the dark matter I-web (yellow dash-dotted line) and the gas I-web (blue dash-dotted line). The upper panel shows the power spectra $P(k)$, while the lower panel shows the ratio between mock and reference power spectrum $P_{\rm mock}(k)/P_{\rm ref}(k)$ for each of the considered bias model. Gray shaded areas stand respectively, from the darker to the lighter, for $1\%, 2\%, 5\%$ residuals.

Mock	$R(\%)$	$\sigma(\%)$
Local $\delta_{\rm dm}$	0.83	0.94
DM I-web	0.78	0.85
Local $\delta_{\rm gas}$	0.67	1.02
Gas I-web	0.58	0.79

Table 5: Percentage residuals R and standard deviation σ of ratios $P_{\rm mock}(k)/P_{\rm mock}(k)$ for the mocks obtained with the considered bias models.

Bispectrum Figure 22 shows reduced bispectrum of the HI mocks, compared to the reference bispectrum, for different configurations in Fourier space and for different tested bias models. The first and second row show respectively reduced bispectra $Q(\theta_{12})$ and ratios $Q_{\rm mock}(\theta_{12})/Q_{\rm ref}(\theta_{12})$, where shaded areas represent error bars. The third and fourth rows represent $\sigma_{\rm ref}$ and $\sigma_{\rm web}$ statistical significance, i.e. respectively the statistical significance of bispectrum obtained with a X cosmic web bias model with respect to the reference and the statistical significance of bispectrum obtained with a a X cosmic web bias model with respect to the one obtained with a Y cosmic web bias model. See equations 50 and 51 (without modulus, as in section 2.1) for the formulas. See Appendix B for a full-page version of Figure 22.

The results reveal that the most accurate reproduction of 3-point point statistics of HI is achieved by adopting the gas I-web model. Less accurate but still good results are obtained with the dark matter I-web. Moreover, the mappings relying only on local overdensities show that the ionized gas carries locally more information on the spatial distribution of HI than dark matter.

Quantitatively, the results achieved including the gas I-web dependence in the bias are compatible within error bars with the reference bispectrum and the statistical significance of such bias model is within $1\,\sigma$ at all scales. Looking at the fourth row of Figure 22, the black dashed and gray dash-dotted line represent respectively the significance of gas I-web model with respect to local DM δ and local gas δ models. We detect a $2-3\,\sigma$ significance of gas I-web with respect to DM δ at all scales and a $\sim 1.5\,\sigma$ significance of gas I-web with respect to gas δ. We regard this as an evidence for long-range non-local bias, independently on whether we use the DM or gas density field as base field. Moreover the purple dotted and brown dash-dotted lines quantifies the significance of respectively gas I-web with respect to DM I-web and local gas δ with respect to local DM δ. In other words, such lines quantify the net gain of information, local and non-local, we achieve by mapping the HI number density field on the ionized gas field, instead of mapping it on the DM density field. This results in a tiny non-local (within $1\,\sigma$) gain of information, while the local information gain achieved by the gas overdensity is $\sim 1.5-2\,\sigma$.

This result is interpreted as follows. As pointed out by Villaescusa-Navarro et al. (2018), HI at $z < 5$ (our realization at $z = 2$ is included in this z range) is mostly contained in dark matter halos. In chapter 4 we have shown that on large scales the dark matter I-web fully determines the spatial distribution and clustering of dark matter halos. Let us suppose that HI is really contained exclusively in dark matter halos. Then, the mappings between dark matter density and respectively halos and HI number density should be similar and the DM I-web is expected to have a significant impact also in the case of HI. In particular, given the high DM-gas cross-correlation on large scales, we believe that the I-webs computed from such fields are

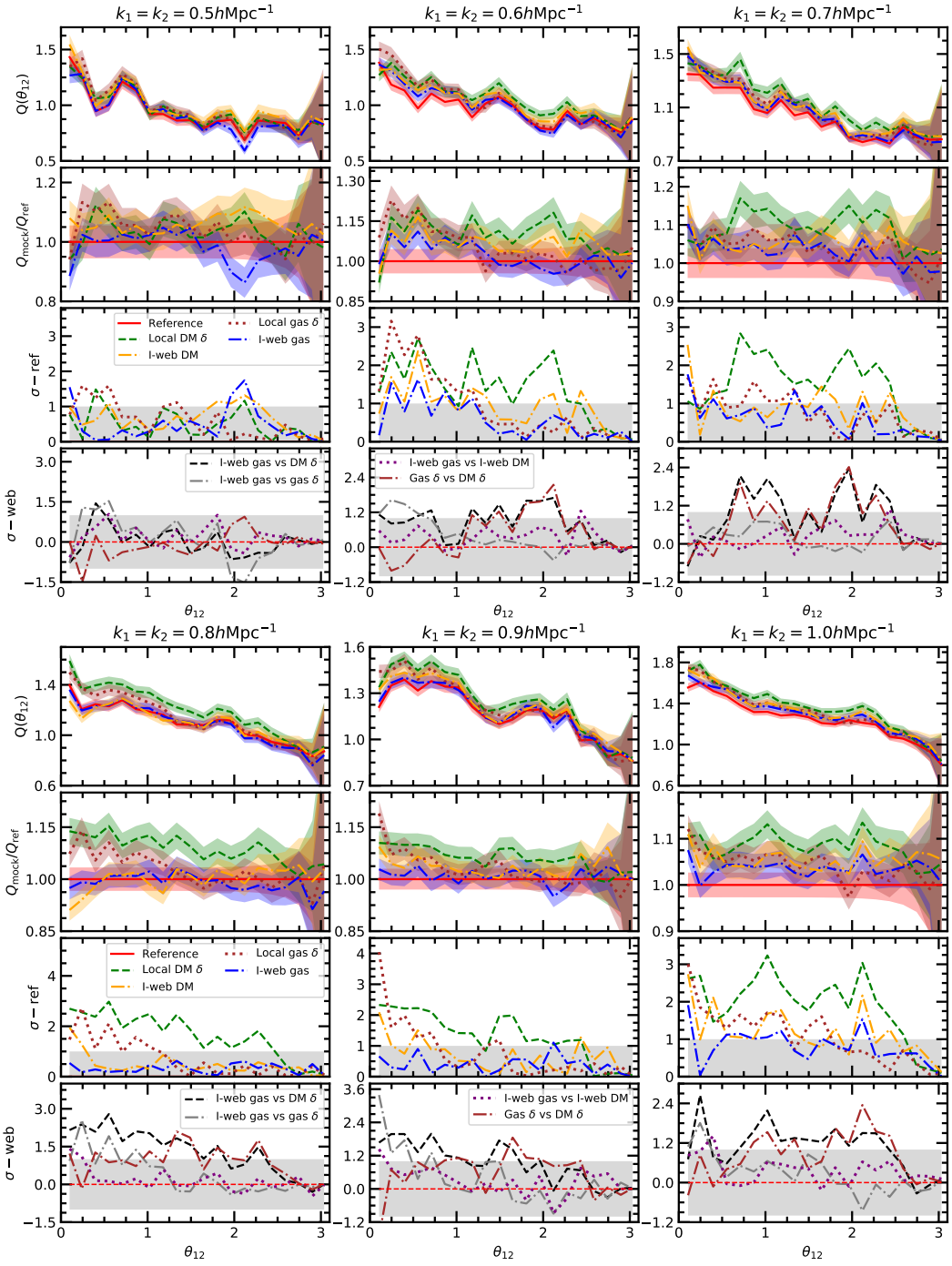

Figure 22: Reduced bispectra of reference field (red solid line) and HI mock catalogs, including in the bias respectively the dependence on local DM overdensity (green dashed line), DM I-web (yellow dashed line), local gas local overdensity (brown dotted line), gas I-web (blue dash-dotted line). In the first row, reduced bispectra $Q(\theta_{12})$ and in the second row, ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$. Shaded areas represent error bars. In the third row and and fourth rows respectively the statistical significance $\sigma - \mathrm{ref}$ with respect to the reference and $\sigma - \mathrm{web}$ of a X bias model with respect to a Y one. Gray shaded areas stand for 1σ significance. The most accurate reproduction of the reference bispectrum is obtained with gas I-web.

practically equivalent and both may fully constrain the large scale spatial distribution of HI. Unfortunately we cannot probe these hypotheses. The typical scales investigated in chapter 4, $k_1 = 0.1, 1, k_2 = 0.2\,h\mathrm{Mpc}^{-1}$ are out of reach in this case. Investigate here scales $k < 0.5h\mathrm{Mpc}^{-1}$ is not very meaningful, given the much smaller volume ($1500h^{-1}$Mpc of `Minerva` versus $100h^{-1}$Mpc of the `Osaka` hydrodynamical simulation). Already at scales $k_1 = k_2 = 0.5h\mathrm{Mpc}^{-1}$ the bispectra look noisy and it is hard to assess which bias model produces the most accurate result, among the ones considered. On smaller scales, the correlation between dark matter and gas density fields slighlty decreases. This means that DM and gas I-webs will no longer be so equivalent as on large scales. On one hand, in chapter 4 we have suggested that on intermediate and small scales the DM I-web could no longer be sufficient to constrain the 3-dimensional spatial distribution of halos. In fact, under the assumption that HI is exclusively found inside DM halos, at the scales investigated here the DM I-web model does not achieve the same accuracy accomplished for halos on the aforementioned large scales. On other hand, the gas I-web models with very high accuracy the 3-point statistics of HI in our analysis.

One possible explanation for such result is that, unlike dark matter, ionized gas and HI interact in baryonic processes on intermediate and small scales, like e.g. photo-heating, photo-ionization and radiative cooling. Therefore, it is not surprising that the spatial distribution of HI is more sensitive to the spatial distribution of gas, rather than of dark matter, which loses track of processes involving baryons on small scales. This argument is quantitatively supported by the fact that locally the gas carries more information on the 3-point statistics of HI than DM. Hence, the gas I-web may account, at least partially, for the effect that such complex non-linear processes have on the spatial distribution of gas.

We conclude that the mapping of HI onto the gas overdensity and its I-web allows to produce accurate mocks of HI number density. This is in summary due to the fact that:

- the local gas density contrains the spatial distribution of HI better than dark matter, due to its baryonic nature;
- the gas I-web is roughly equivalent to the dark matter I-web on large scales due to the $\sim$ 100% DM-gas cross-correlation, but on smaller scales it may be more sensitive than DM I-web to the baryonic processes which involve gas and HI.

These findings are useful also in the perspective of creating mock 21cm-line maps. Moreover, as argued by Villaescusa-Navarro et al. (2018), the fact that the HI is mainly contained in halos opens the possibility of creating HI mock catalogs just by running N-body simulations and then properly populating dark matter halos with HI, instead of running full hydrodynamical simulations.

Differently from HI, the ionized gas is not localized mainly inside dark matter halos, but a large fraction of it is found in the IGM outside collapsed objects. We regard this difference is the main reason why including DM I-web in the mapping of gas onto the dark matter field does not have the same significant impact it has on the HI mapping.

Reasoning again in the PT perspective of McDonald & Roy (2009), given the results we have obtained one could think of formalizing a PT expansion which expresses the HI overdensity as a function not of DM overdensity but of ionized gas density and to build local and non-local terms from the latter. The advantage of relying on gas is that on large scales, where standard PT holds, the DM-gas cross-correlation is $C(k) \sim 100\%$ and perturbative expansions built separately on DM and on gas would be basically equivalent. However, if one could consider the full non-linear density fields and go down to small scales, the gas density field would describe better the spatial distribution of HI. Therefore, we propose to consider defining a bias expansion for HI as a function of the gas density.

2.3 Temperature

The temperature requires a more complicated treatment. Let us start by simply mapping the temperature onto the DM field. According to Figure 15, such fields have a low cross-correlations on small scales. As already anticipated, when this happens the bias functional dependence between DM and temperature is not physically correct and `BAM` does not converge properly to the reference power spectrum. This is due to the fact that the bias constrains the assignment of the tracer field (temperature, in this case) in an incorrect way. In this case the code converges, but the final mock power spectrum does not reproduce the reference power spectrum within the required accuracy. Figure 26 shows the initial and final steps of the convergence of the mock power spectrum. The red solid line represents the reference power spectrum, the green dashed line the mock power spectrum at the first iteration performed by the code (Iteration 0), the blue dash-dotted line the mock power spectrum the maximum reached convergence, at Iteration 200. The first mock (Iteration 0) loses much power on large scales with respect to the reference field ($R \sim 500\%$. In a normal context, $R \sim 20 - 30\%$ at Iteration 0. When the kernel is applied, the iterative process is able to make the $P(k)$ converge only to $R \sim 50\%$ residuals in the power spectrum and gets stuck at that point. This makes it clear that proceeding in this way does not lead to the required convergence.

One could wonder whether acting in the same spirit of the HI and using a mapping onto one of the other baryonic fields, i.e. gas density or HI number density, can overcome this problem. However, an explicit check we have performed shows that none of the other fields guarantees the proper convergence. Instead, they provide an equally bad or even worse convergence

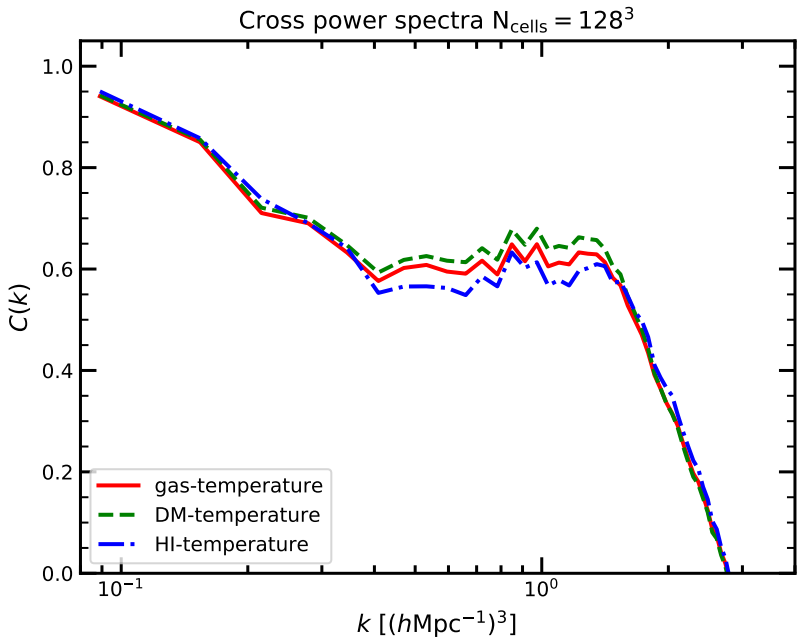

Figure 23: Normalized cross power spectrum $C(k)$ of temperature and respectively DM density (green dashed line), gas density (red solid line) and HI number density (blue dash-dotted line) fields.

scenario. This can be understood by looking at Figure 23, which shows normalized cross power spectrum $C(k)$ of temperature and respectively DM density, gas density and HI number density. One clearly notices that such cross-correlations are similar one another and are all equally low. Hence, the failure of such mapping should not come as a surprise. Therefore, an alternative treatment is needed here. As anticipated previously, the temperature is determined by the photo-ionization equilibrium condition, described by equation 38 in chapter 2.

Let us consider for a moment that ionized gas consists only of ionized H (with number density n_{HII}). Because $n_e\, n_P \approx n_{\mathrm{HII}}^2$, knowing the temperature and the analytical expression for the UV photon background would mean constraining the ratio $n_{\mathrm{HII}}/n_{\mathrm{HI}}$.

Here, none of the two assumptions hold exactly, i.e. neither the ionized gas is all ionized H, nor we know the UV photon background. However, notice that the reference simulation basically considers H and He as baryonic species. Let us also consider that the photo-ionization of He is a much more rare event, which requires more energetic photons. Moreover, almost all the He is ionized after He Reionization and neutral He is much less abundant than neutral H. As a result, if we neglect neutral He and group all the ionized gas species together (with number density n_{gas}), the ratio $n_{\mathrm{gas}}/n_{\mathrm{HI}} \sim n_{\mathrm{HII}}/n_{\mathrm{HI}}$ partially constrains the temperature.

Therefore, the idea we apply here to overcome the convergence problem is to exploit the photo-ionization equilibrium. This is done by mapping the temperature onto the dark matter field, to which the kernel is applied, but explicitly including the ionized gas density and HI number density as additional properties in the bias. The purpose of this operation is introducing in the bias the information on the relative abundance between ionized gas and neutral hydrogen (or equivalently, between ionized and neutral chemical species) and constrain the temperature.

Even though a full auto-consistent convergence is still not reached, this simple procedure allows to reduce mock power spectrum average residuals from $R \sim 50\%$ to $R \sim 5\%$. Then, with an additional post-processing operation we manage to further reduce the residuals to $R \sim 1\%$. We start by noticing that the mock power spectrum is characterized by a $\sim 5\%$ nearly-constant bias with respect to the reference power spectrum. In fact, the mock $P(k)$ lies systematically below the reference $P(k)$ by $\sim 5\%$, expect for a little further deviation on intermediate scales. To fix such problem, we apply a convolution with a constant kernel (i.e. constant at all scales), which has the effect of uniformly increasing the power in all the spherical shells in Fourier space and artificially make the mock $P(k)$ fit the reference $P(k)$. The details of the full procedure are described in what follows.

Qualitative assessment

To begin with, as done for the other components of the IGM, we first perform a visual comparison of slices of the catalogs. Figure 24 shows slices of the reference simulation (upper left) and of the mocks obtained respectively with the mapping only on the DM field (upper right), including in the bias gas and HI (lower left), including in the bias gas and HI and performing the additional convolution (lower right). One immediately notices that the mock obtained with the mapping only on the DM field does not resemble at all the reference field. Instead, both the mocks obtained accounting for gas and HI, both with and without convolution, look very similar to the reference temperature field. Already at this point we can appreciate how the reproduction of the temperature map significantly improves when a proper bias relation is adopted, in the spirit of fulfilling the photo-ionization equilibrium condition.

Moreover, we notice that the mocks obtained including gas and HI in the bias look very similar, which means that the convolution by a constant factor does not spoil the original mock.

Quantitative assessment

Summary statistics: power spectrum In this section we analyze the results concerning the power spectrum of temperature.

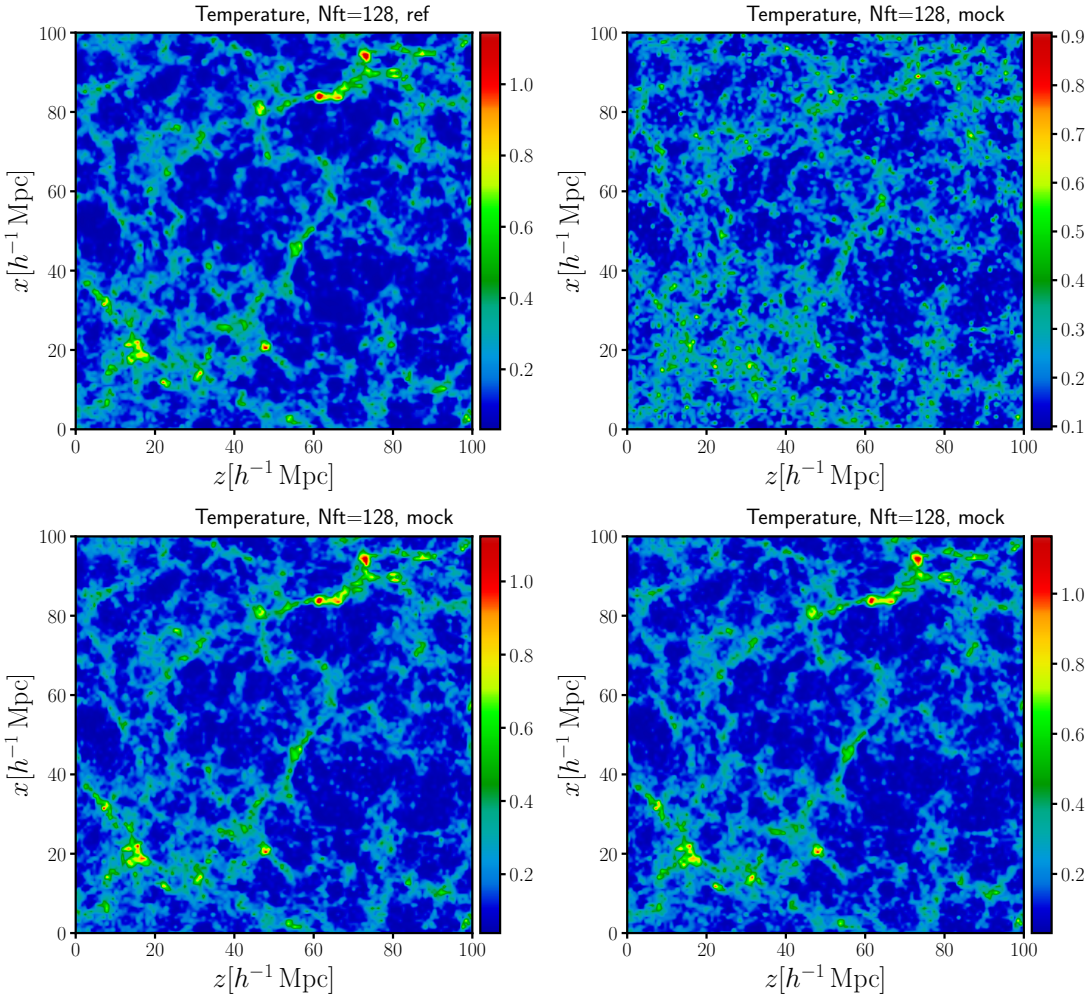

Figure 24: Slices of the temperature field, extracted at the same position of the mesh. In the upper left panel, a slice of the reference simulation box. In the upper right panel, a slice of the mock produced with the simple mapping on the DM field. In the lower panels, respectively the mock obtained introducing the dependence on gas and HI fields in the bias (lower left) and the same mock + the additional (lower right). While the upper right mock does not reproduce the visual aspect of the reference, the lower panels do. Hence, this points towards the fact that fulfilling the photo-ionization equilibrium condition helps to accurately reproduce the temperature field. The lower panels differs by a convolution for a constant kernel in Fourier space, i.e. by a constant in configuration space, which is re-absorbed when computing the temperature overdensity.

As already shown in Figure 26, a simple local DM bias model does not lead to a proper convergence. Hence, we discuss here directly the results regarding the case in which we include in the bias the dependence on gas density and HI number density.

Figure 25 shows the comparison between the reference (red solid line) and the mock power spectra obtained including in the stochastic bias respectively the local gas and HI overdensities (green dashed line) and the same model + the additional convolution (blue dash-dotted line). The upper panel shows the power spectra $P(k)$, while the lower panel shows the ratio between mock and reference power spectrum $P_{\rm mock}(k)/P_{\rm ref}(k)$ for each of the considered bias model. Gray shaded areas stand respectively, from the darker to the lighter, for $1\%, 2\%, 5\%$ residuals. From a visual comparison it turns out that in the case without convolution the power spectrum is not accurate, because is deviates systematically by $\sim 5\%$ from the reference power spectrum. However, we notice that the deviation is quite uniform, i.e. it remains nearly constant in all the spherical shells in Fourier space, expect for a little further deviation on intermediate scales. Moreover, Figure 27 shows that this behaviour does not depend significantly on iterations. Once the convergence is achieved, the $P(k)$ remains constant (beside fluctuations) along the iterative process. We regard that such calibrated bias is physically correct. The converged mock power spectrum has indeed the correct shape, but has uniformly wrong amplitude at all scales.

In this perspective, we decide to adopt the aforementioned post-processing operation, consisting in a convolution with a constant kernel in all the spherical shells in Fourier space. Namely, the Fourier-transformed contrast temperature field is multiplied by a constant. This shifts the $P(k)$ and makes it fit artificially the reference $P(k)$. The constant is chosen in such a way to minimize the residuals. The blue solid line in Figure 25 shows the result of this operation, with kernel $\mathcal{K} = 1.005\,\forall k$, which results in $R = 1.01$ average residuals.

One could argue that such post-processing operation is in principle a bit arbitrary because of the choice of the multiplying factor in Fourier space. This is general true and cannot be avoided. In this case, in which we can rely on the reference temperature field, we can tune the factor in such a way to minimize the average residuals R. However, in a general context of mock production, the temperature field is built starting from the DM field obtained with ALPT and following the mapping procedure presented at the beginning of the chapter. Hence, in that case one does not have the temperature reference field and cannot choose the multiplying factor as we have done in this case. However, let us consider the two following argument:

- in the case of a standard convergence, we expect the calibrated bias relation to produce mock catalogs (using a DM field different from the one used for the calibration) whose power spectrum reproduces

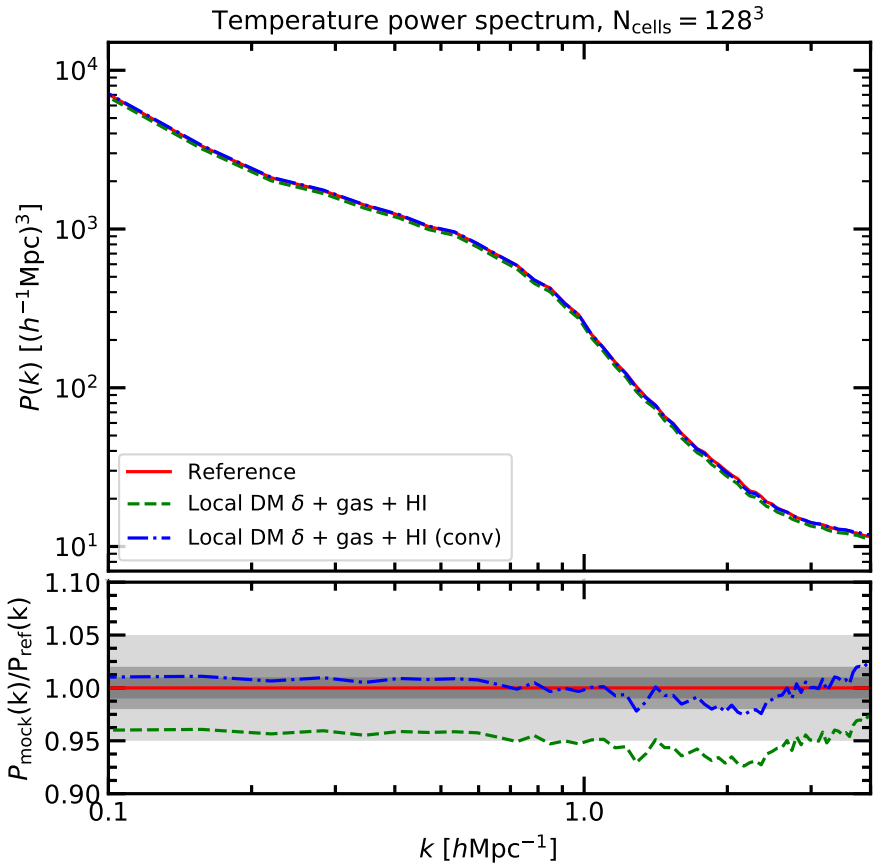

Figure 25: Comparison between reference (red solid line) and temperature mock power spectra, the latter obtained including in the bias the dependence on the local gas and HI overdensities (green dashed line) and the same model + the additional convolution with $\mathcal{K} = \text{const}, \forall k$. The upper panel shows the power spectra $P(k)$, while the lower panel shows the ratio between mock and reference power spectrum $P_{\text{mock}}(k)/P_{\text{ref}}(k)$ for each of the considered bias model. Gray shaded areas stand respectively, from the darker to the lighter, for $1\%, 2\%, 5\%$ residuals.

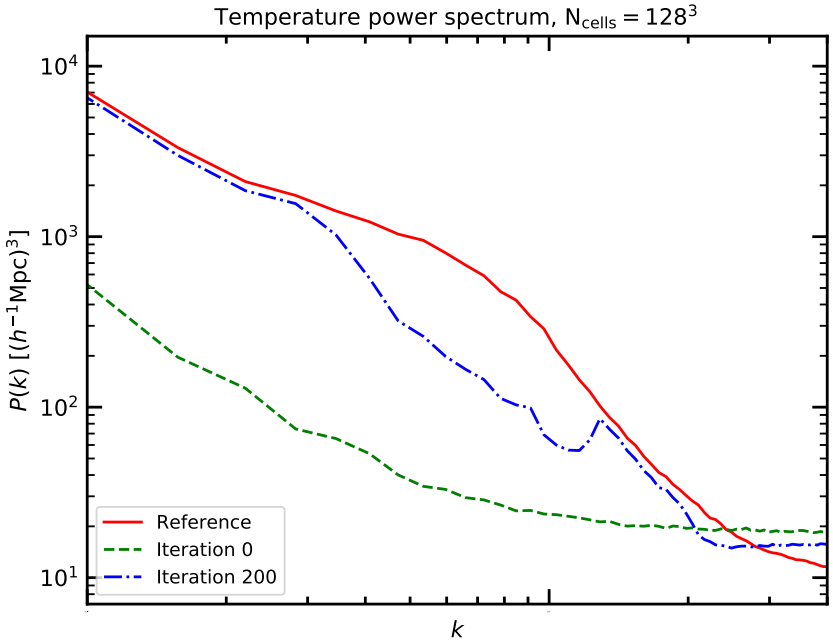

Figure 26: Power spectra of the reference temperature field (red solid line) and mock catalogs obtained including only DM δ in the bias, at Iteration 0 (green dashed line), i.e. the first mock created by BAM, and at Iteration 200 (blue dash-dotted line), at the maximum convergence. The plot clearly shows that this bias model is not able to drive the convergence in such a way that the mock $P(k)$ ends on top of the reference $P(k)$.

the reference power spectrum with high accuracy. Equivalently, here we believe we can regard that the calibrated bias relation produces temperature mocks which systematically underestimates by 5% the true power spectrum, in a nearly constant way at all scales;

- it may happen that the multiplying factor which minimizes the residuals is not exactly the same for all the mocks. This could due e.g. on the intrinsic shape of the power spectrum or on some fluctuations. However, the $P(k)$ of the same tracer field of different realizations tends to converge to the same amplitude at small scales. Hence, the constant kernel $\mathcal{K}$ can be at first applied to mocks of different realizations, check whether the resulting mock $P(k)$ reproduces the amplitude of the reference $P(k)$ on small scales and, if not, re-tune the constant kernel $\mathcal{K}$ in such a way to match the amplitude of the reference power spectrum.

In this way, the procedure still suffers some arbitrariness, but we believe these arguments helps to constrain the choice of the constant kernel. For these reasons, we regard this operation can be considered as a post-processing part of the calibration, rather than an arbitrary multiplication by a constant factor.

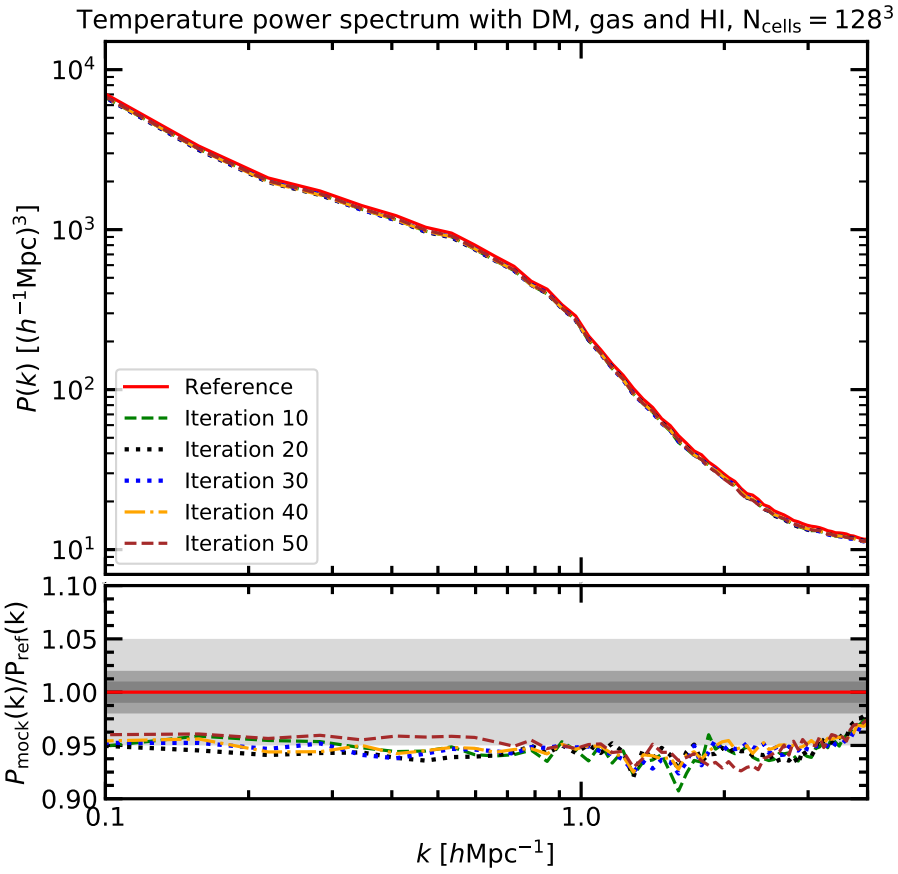

Figure 27: Mock temperature power spectrum obtained with the mapping on DM, gas and HI, compared to the reference power spectrum (red solid line). The mock power spectra are characterized by a $\sim$ 5% nearly constant bias with respect to the reference power spectrum in all the spherical shells in Fourier space, except for a little further deviation on intermediate scales. Such behaviour remains stable along the iterations, once the convergence has been achieved (here already at Iteration 10).

Anyway, this procedure is still sub-optimal. Future investigation will try to find an alternative way to map the temperature, in such a way to achieve an automatic mapping which does not need the additional convolution.

In the following paragraph, we will see that using DM, gas and HI fields simultaneously in the bias and the additional convolution leads to an extremely good accuracy in the 3-point statistics.

Summary statistics: bispectrum Figure 28 shows reduced bispectrum of the temperature mocks, compared to the reference bispectrum, for different configurations in Fourier space and for different tested bias models. The first and second row show respectively reduced bispectra $Q(\theta_{12})$ and ratios $Q_{\rm mock}(\theta_{12})/Q_{\rm ref}(\theta_{12})$, where shaded areas represent error bars. The third row represents $\sigma_{\rm ref}$ statistical significance, i.e. the statistical significance of bispectrum obtained with a X cosmic web bias model with respect to the reference. See equations 50 for the used formula. See Appendix B for a full-page version of Figure 28. Both the studied cases reproduce very accurately the reference bispectrum. The mock bispectra are both within $1\,\sigma$ significance almost at all scales. This fact supports that the stochastic bias model we are investigating is physically well-motivated. Moreover, the best result is achieved with the convolution: the resulting bispectrum results in a $\sigma-$ref significance which is within $0.5\,\sigma$ at all scales, except for the configuration $k_1 = k_2 = 1.0h\text{Mpc}^{-1}$, at which anyway is within $1\,\sigma$. Therefore, the convolution not only forces the power spectrum to have the correct amplitude, but also reproduces the 3-point statistics with great accuracy.

This way of proceeding is interesting because we start from 3 fields with a low cross-correlation with the temperature on small scales. When considered separately, the functional dependence between the temperature and each of them is thought to be physically correct on large scales, while is not on small scales. However, when all the 3 fields are simultaneously taken into account, we manage to reproduce very accurately the reference temperature field. Although such bias relation can still be refined, we believe that embedding in the mapping the dependence on different fields at the same time is an important aspect.

This result can be understood from two perspectives:

- from a phenomenological point of view, we are including in the bias some information about the physical processes which involve the baryonic components of the IGM and determine the temperature;
- from a PT point of view, one can think of formalizing an expansion which does not express the temperature as a function of local and non-local terms computed from just one field, but computed instead from all the 3 fields, in a multi-field fashion. Here, we have investigated a bias model restricted to the local dependence on DM, gas and HI.

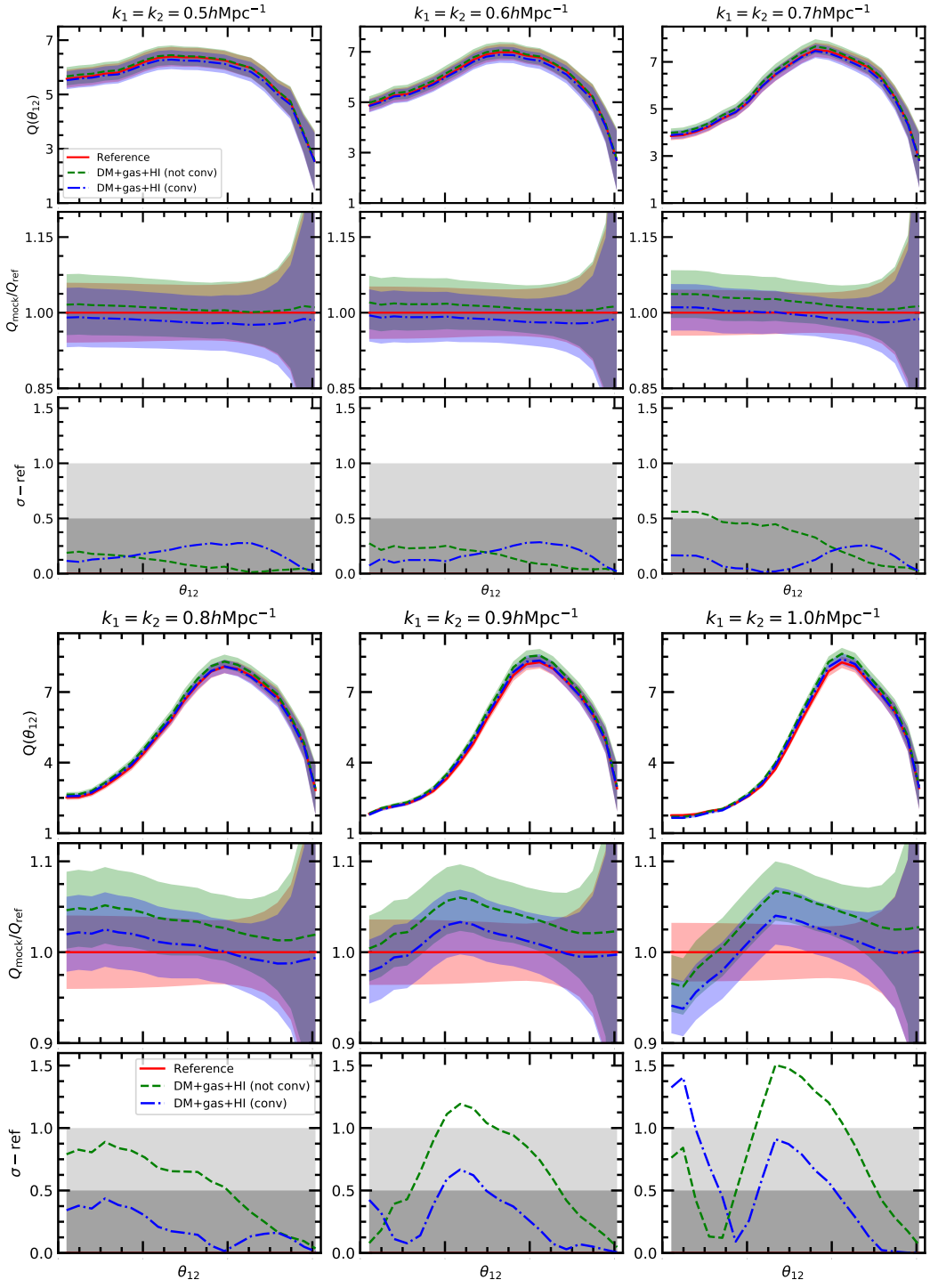

Figure 28: Reduced bispectra of the reference field (red solid line) and temperature mock catalogs, including in the bias respectively the dependence on gas and HI overdensities (green dashed line) and the same + the convolution by $\mathcal{K} = \text{const}$, $\forall k$ (blue dash-dotted line). In the first row, reduced bispectra $Q(\theta_{12})$ and in the second row, ratios $Q_{\text{mock}}(\theta_{12})/Q_{\text{ref}}(\theta_{12})$. Shaded areas represent error bars. In the third row the statistical significance $\sigma - \text{ref}$ with respect to the reference. Gray shaded areas stand for 1σ (lighter) and 0.5σ (darker) significance. The most accurate reproduction of the reference bispectrum is obtained by accounting for gas and HI overdensities in the bias and applying the further convolution.

Such model, with the aid of a proper post-processing convolution, is found to accurately reproduce the 2- and 3- point statistics at the considered scales. However, the possibility that on other scales we could need additional non-local dependencies is not to be ruled out.

2.4 Optical depths

The final quantity we want to be able to map is the optical depth for Lyman-α photons. As already anticipated, the optical depth of the IGM is commonly computed in these studies with the Fluctuating Gunn-Peterson Approximation (FGPA), described by equation 40 in chapter 2, both from a theoretical point of view (see Cieplak & Slosar, 2016; Gunn & Peterson, 1965; Seljak, 2012; Weinberg et al., 1997) and for the purpose of producing Lyman-α mock catalogs (e.g. Farr et al., 2020; Porqueres et al., 2019). This approximation is assumed because it allows to rapidly compute the optical depth field from the matter density field just by taking a non-linear transformation of the latter.

However, the formal definition of optical depth at a given frequency is

$$\tau_\nu = \int_{s_1}^{s_2} k_\nu \, \rho \, \mathrm{d}s \quad , \tag{57}$$

where k_ν is the opacity at frequency ν, ρ is the matter density responsible for the opacity and the integral is performed along the line-of-sight. That is to say, in order to compute τ exactly one has to perform an integration along the all the possible lines-of-sight of the simulation box. Concretely, such integral is discretized and computed in each cell x as a proper sum (Nagamine et al., 2020)

$$\tau(x) = \frac{\pi e^2}{m_e c} \sum_j f \, \phi(x - x_j) \, n_{\mathrm{HI}}(x_j) \, \mathrm{d}l \quad , \tag{58}$$

where e is the electron charge, m_e is the electron mass, c is the speed of light, f is the oscillator strenght, n_{HI} is the HI number density, x_i is the location of the jth-cell and ϕ is the Voigt profile computed with the fitting formula by Tasitsiomi (2006) without direct integration.

Because the FGPA approximation is supposed to be good for the optical depth, at least on large scales, many studies rely on that because it is much simpler than performing the full integration.

However, as anticipated previously, the optical depth field computed exactly (τ hereafter) has been kindly provided by the Theoretical Astrophysics research group of Osaka University.

The outline of this section consists in comparing τ with the optical depth computed with the FGPA approximation (τ_{FGPA} hereafter) and trying to

calibrate the bias relation in both ways, in order to sample mock catalogs of optical depths and Lyman-α forest fluxes. Our preferential scenario will be being able to sample accurate mocks of the exact τ. The FGPA approximation is calculated with equation 40, using $\alpha = 1.6$ and $A = 0.17$. The value of the constant A is computed using the operative definition by Seljak (2012)

$$A(z) = 0.0012\,(1+z)^{9/2} \quad . \tag{59}$$

At the resolution we are considering, one could think of producing Lyman-alpha forest mocks in the following way:

- start by producing one HI mock;
- compute the exact optical depth respectively from HI number density mocks with a proper integration along the line-of-sight;
- compute the transmitted fluxes with non-linear transformation $F = \exp(-\tau)$.

Because the HI mocks reproduce the reference simulation very accurately, we may expect to obtain in this way Lyman-α forest mock catalogs which preserve the correct spatial correlations.

However, in order to save time and calculations and try to extract physical information on the Lyman-α forest bias, we explore here whether a more direct mapping can be used. Thus, we study how to map directly the optical depth onto the DM field.

Let us now try to understand which are the differences between the τ and $\tau_{\rm FGPA}$ fields at the level of realization. Figure 29 shows a comparison between slices: in the top row optical depths τ (on the left) and $\tau_{\rm FGPA}$ (on the right), in the bottom row Lyman-α fluxes, computed respectively as $F = \exp(-\tau)$ (on the left) and $F_{\rm FGPA} = \exp(-\tau_{\rm FGPA})$ (on the right). The fluxes F are actually normalized to the continuum un-attenuated flux F_c, so that what the plot really represents is the normalized transmitted flux F/F_c. As in previous cases, for plotting purposes the optical depth slices are converted to their corresponding overdensities δ (or, more properly, to contrast τ) and then to $\log_{10}(2+\delta)$. This is done in order to visually enhance the visual aspect of underdense regions. The flux slices are instead plotted in linear scale. One clearly notices that, at the high resolution we are looking at (the lowest resolution available, which is anyway a very high resolution), τ and $\tau_{\rm FGPA}$ and their corresponding Lyman-α fluxes do look different.

In order to understand in an even clearer way the difference between the exact and FGPA fluxes, we placed skewers along the line-of-sight (i.e., along the direction of integration in the τ field) in both exact FGPA and exact reference flux fields. Figure 30 shows the comparison between such skewers as a function of their position along the line-of-sight. The resulting

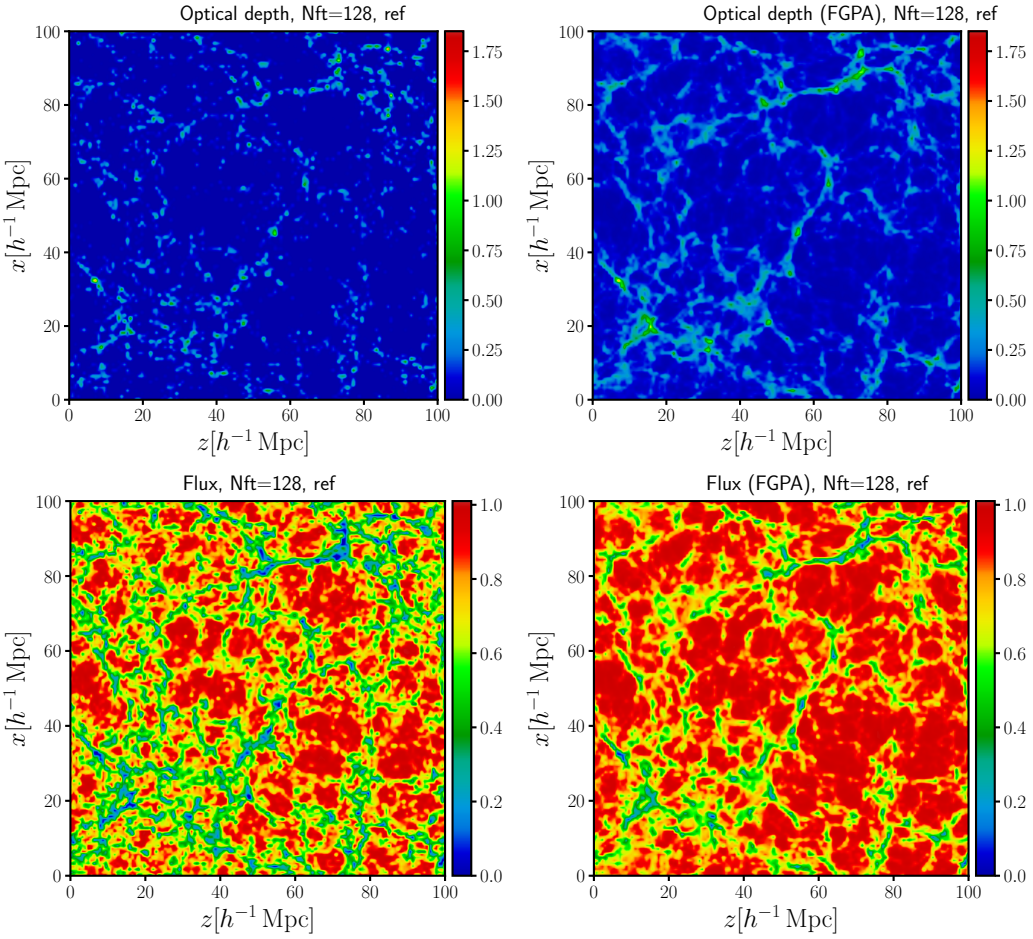

Figure 29: Slices of the reference simulation fields, extracted at the same position of the mesh. In the upper left panel, a slice of the exact τ field, computed exactly by integrating along the line-of-sight; in the upper right panel, a slice of the τ_{FGPA} field, computed with the FGPA approximation; in the lower left panel a slice of the exact Lyman-α fluxes computed as $F = \exp(-\tau)$; in the lower right panel a slice of the FGPA Lyman-α fluxes, computed as $F_{\mathrm{FGPA}} = \exp(-\tau_{\mathrm{FGPA}})$.

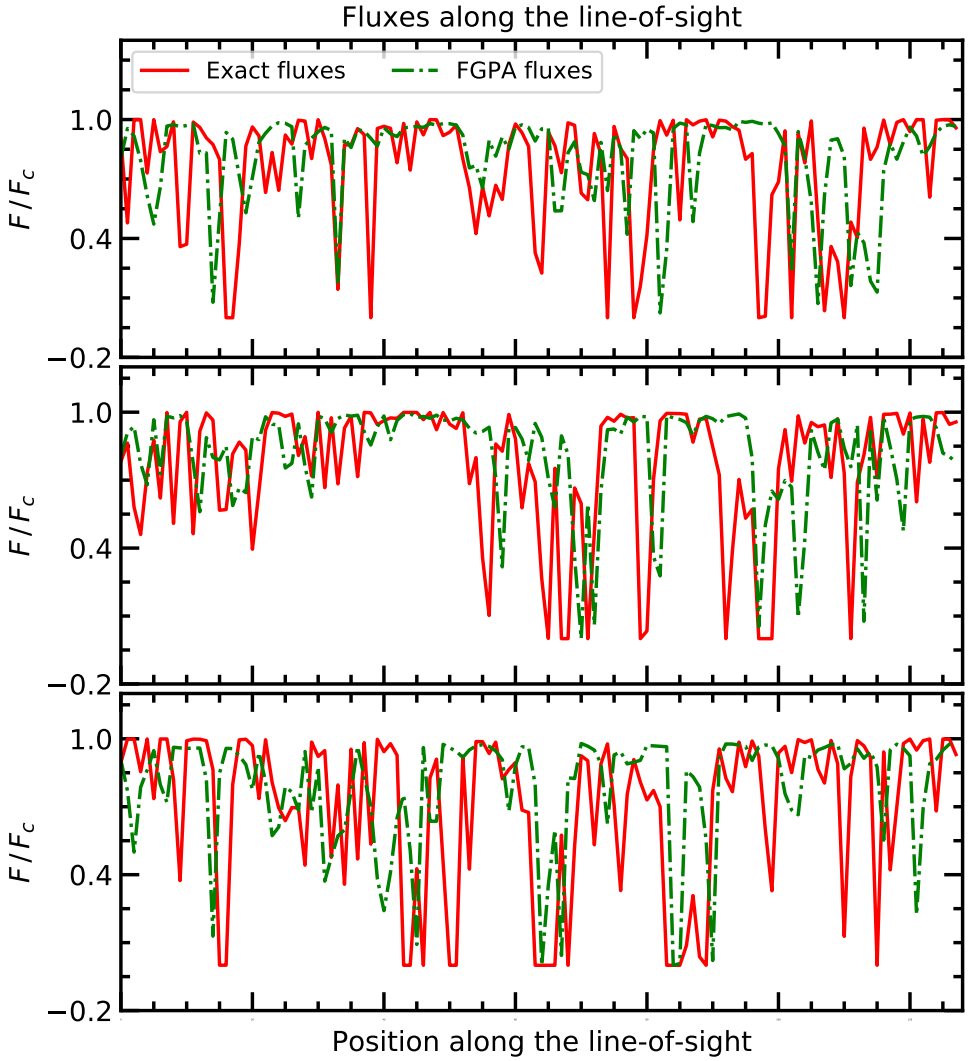

Figure 30: Transmitted flux F/F_c in three different skewers, along the direction of integration on the exact case. The red solid lines represent the exact fluxes, the green dash-dot lines represent the FGPA fluxes.

plots provide an idea of the final aspect the observed Lyman-α forest in real space, normalized to the continuum of the quasar. It turns out that the FGPA approximation produces a weaker absorption effect with respect to the exact integration. This is clear above all in the underdense regions, in which the lines do not saturate. Moreover, a careful observation of the spectra reveals that the most prominent lines in the exact computation are present also in the FGPA approximation, but look slightly shifted. Notice that, when one chooses a lower resolution, the differences at the level of realization between the exact and approximated calculation of the flux fields are likely not to look as evident as in this case.

Finally, we look at the power spectra. Figure 31 shows a comparison between the τ and τ_{FGPA} fields power spectra. One clearly sees that the $P(k)$

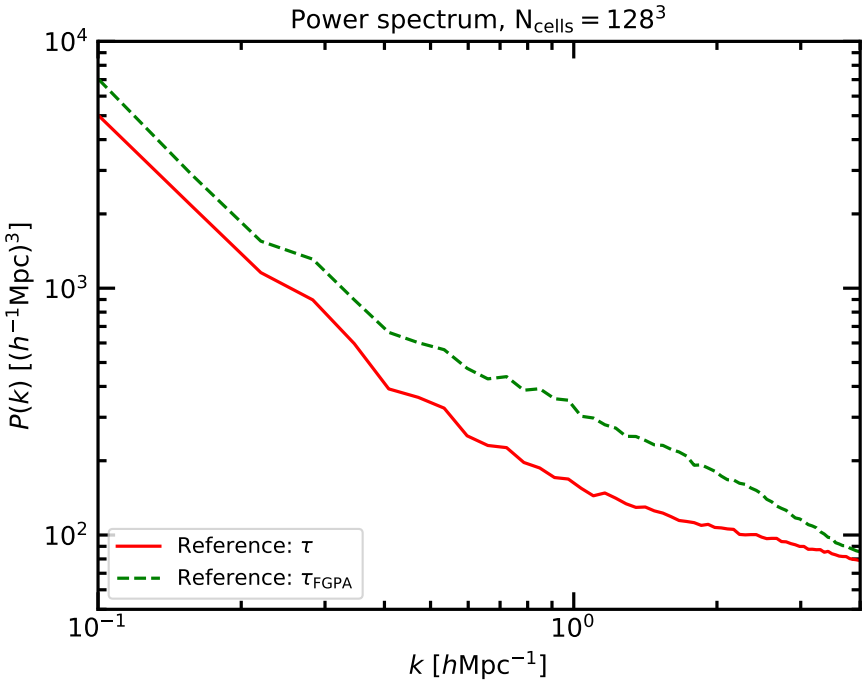

Figure 31: Comparison between power spectra of the optical depth τ, computed either exactly by integrating along the line-of-sight (red solid line) or with the FGPA approximation (green dashed line).

look quite different and that the FGPA approximation gives more power, i.e. larger clustering, at all scales.

This preliminary analysis motivates why it is worth trying to find a mapping of the exact τ field, in such a way to be able to extract correctly the Lyman-α forest spectra.

Because a large number of Lyman-α forest mock catalogs will be needed, without the necessity of producing mocks of the entire hydrodynamical simulations, the first attempt we perform is trying to map τ directly onto the DM field. Figure 32 shows normalized cross-spectra $C(k)$ for a number of couple of fields. The blue dash-dotted line represents the cross-correlation between the DM density field and the HI number density field. One could wonder whether the calculation of the optical depth, either by integration of HI or by the FGPA non-linear transformation of gas, makes it gain or lose cross-correlation with the DM field. The green dashed line, corresponding the to the DM-τ_{FGPA} normalized cross-spectrum, reveals that the cross-correlation keeps being high and hence we do not expect to experience particular problems in handling the mapping of the FGPA approximation. However, the red solid, which represents the DM-τ cross-power, tells us that the integration procedure lowers dramatically the DM-τ cross-correlation on small scales. We hence start by considering that the mapping of the exact τ onto the DM field may be non-trivial and we are likely to encounter

problems of convergence and/or of stability of the code. Moreover, the τ-$\tau_{\rm FGPA}$ cross-correlation (black dotted line) is the lowest among the one considered. This means that the two optical depth fields are significantly different, especially at intermediate and small scales. We regard this is mainly due to the fact that the integration along the line-of-sight introduces an anisotropy, which is not present in the FGPA approximation. The FGPA approximation is shown to represent quite correctly the exact τ field on large scales ($C(k) \sim 95\%$), while on small scales the correlation drops to $C(k) \sim 10\%$. This is quite intuitive if one considers that the integration involves a limited number of cells around the one considered and thus the smaller scales are the ones which are more affected by such anisotropy. These arguments makes it clear the FGPA approximation does not represent properly the optical depth field on small scales, at the resolution we are working. Therefore, taking into account the anisotropy introduced by the integration, or in other words the privileged direction defined by the line-of-sight of the observer, is an aspect which appears to be not negligible.

In order to find a proper mapping of the exact τ, one attempt we perform is that of introducing in the DM field an anisotropy along the line-of-sight. The resulting quantity, dubbed $\tau_{\rm DM}$[1] is obtained by performing an integration of the DM density field along the line-of-sight in a simplified version of the one described by equation 58. The way we do it is by neglecting all the constants, which are re-absorbed in the auto/cross-power computation when passing to overdensity, and employing a Gaussian profile instead of a Voigt profile. I.e., we compute

$$\tau_{\rm DM}(x) = \sum_j \phi(x - x_j)\, n_{\rm DM}\, {\rm d}l \quad , \tag{60}$$

where $\phi(x)$ is a Gaussian profile with zero mean (i.e., centered on the considered cell) and standard deviation σ. The profile is truncated at $x \pm 3\sigma$, i.e. $\phi(x) = 0$ for $|x| > 3\sigma$. This operation is by far faster than the full integration of equation 58 and takes a typical time-scale of the order of few tens of seconds. Therefore, it is a feasible computation in the perspective of a massive production of mock catalogs.
In Figure 32, the yellow dash-dotted line shows the $C(k)$ of $\tau_{\rm DM}$ and τ, where $\tau_{\rm DM}$ is calculated with a Gaussian profile with $\sigma = 5$ (in units of cells). We notice that this operation leaves the cross-power unchanged on large scales, while on small scales we detect a gain of $\sim 20\%$ in $C(k)$ (yellow dash-dotted line vs red solid line).

In what follows we test this idea in the code in order to try to map the exact τ. We will observe that using $\tau_{\rm DM}$ instead of DM density improves

[1] Because DM is supposed not to interact with radiation, in principle defining an optical depth for DM does not make sense from a physical point of view. However, we adopt such denomination to indicate that the DM has been modified to mimic the anisotropy present in the exact τ field.

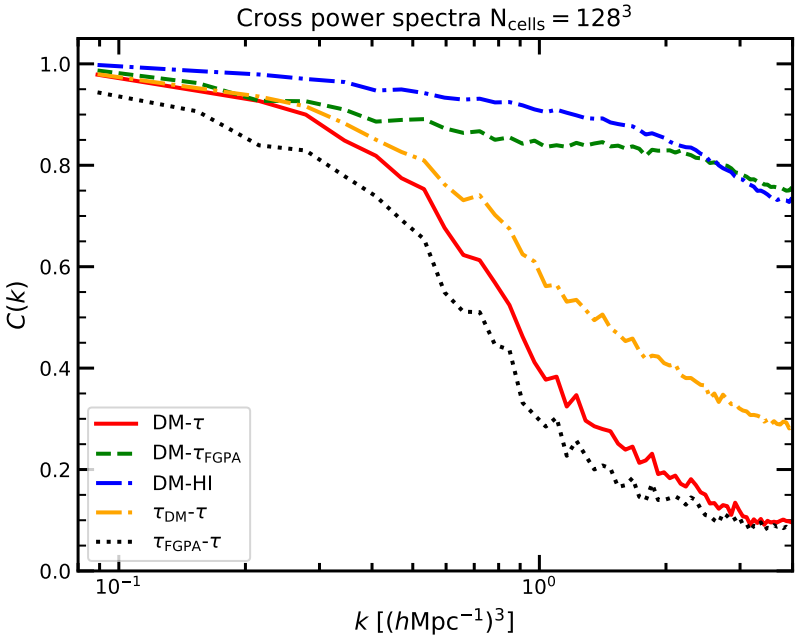

Figure 32: Normalized cross-power spectra $C(k)$ of DM and τ (red solid line), DM and τ_{FGPA} (green dashed line), DM and HI (blue dash-dotted line), properly-defined DM optical depth τ_{DM} and τ (yellow dash-dotted line), τ_{FGPA} and τ (black dotted line). One clearly notices how the FGPA approximation produces an optical depth field which has a low cross-correlation with the τ field computed exactly

the mapping of τ. However, the result are still not at the required level of accuracy.

Finally, we test also the mapping of the $\tau_{\rm FGPA}$ optical depth, in order to show that the BAM method is anyway able to handle the FGPA approximation.

Mapping the exact optical depth

In this section we explore how to map the exact τ field first onto the DM field and then onto the $\tau_{\rm DM}$ field. As already anticipated, the mapping onto the DM field is expected to be unsuccessful. Though, our expectation is that using the $\tau_{\rm DM}$ field instead of the DM field may improve the mapping.
Figure 33 shows a comparison between the reference power spectrum of τ with the power spectra of the optical depth mock catalog obtained respectively with the mapping on the local DM density (green dashed line) and on the DM optical depth $\tau_{\rm DM}$ (blue dash-dotted line). The upper panel shows the power spectra, while the lower panel shows the ratio $P_{\rm mock}(k)/P_{\rm ref}(k)$ between mock and reference power spectra. The case of the mapping on DM density has been taken out from the lower panel for plotting reasons. It is already very clear from the upper panel that the bias model based to $\tau_{\rm DM}$ improves the one based on DM density and allows to achieve a better convergence of the mock power spectrum.

The target power spectrum convergence has not been achieved yet. However, the average residuals R of the mock power spectrum with respect to the reference power spectrum are lowered from $R \sim 27\%$ to $R \sim 10\%$ when τ is mapped onto $\tau_{\rm DM}$ rather than onto the DM density. Therefore, even though more work needs to be done, we believe this preliminary operation we have perforemd on the DM field in order to introduce an anisotropy goes in a promising direction. One possible way to further exploit this idea is applying the simplified integration presented in equation 60 to the HI number density field in order to compute a simplified version of the optical depth onto which map the exact optical depth.

Once the power spectrum convergence will be solved and a convincing well-motivated mapping will have been found, one can repeat the same analysis performed in the previous sections in order to understand which dependencies are to be included in the stochastic bias in order to accurately reproduce the 3-point statistics. We do not look at the bispectrum here, because a mandatory requirement is to have a well-converged 2-point statistics, at least within 2-3% with respect to the reference simulation, before studying higher order statistics.

We leave these questions as an open issue, to be addressed in the near future.

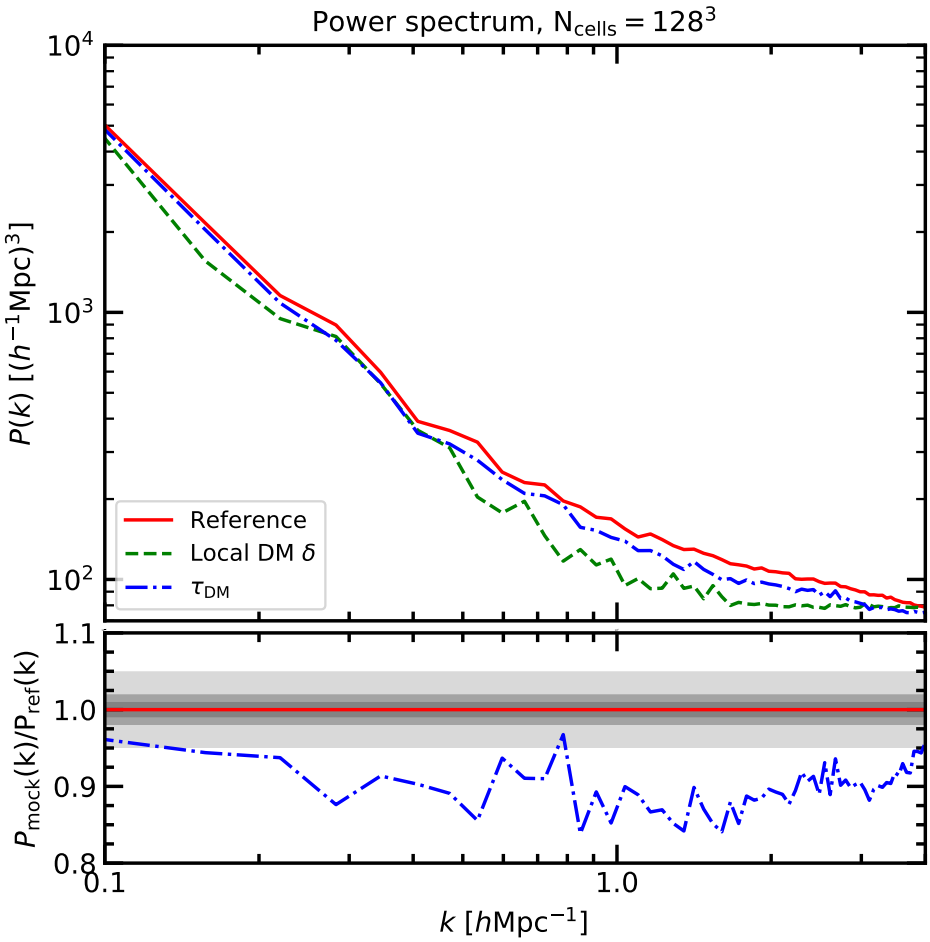

Figure 33: Comparison between the power spectrum of the reference τ field with the power spectra of the optical depth mock catalog obtained by including in the bias the dependence on the local DM density (green dashed line) and the fictious DM optical depth τ_{DM} (blue dash-dotted line), the latter obtained as described by equation 60. The upper panel shows the power spectra, while the lower panel shows the ratio $P_{mock}(k)/P_{ref}(k)$ between mock and reference power spectra. The DM density case has been taken out in the lower panel because it would have forced to enlarge too much the range in the vertical axis. Anyway, the upper panel is in this case already sufficient to conclude that using τ_{DM} instead of DM density allows to improve the convergence of the power spectrum.

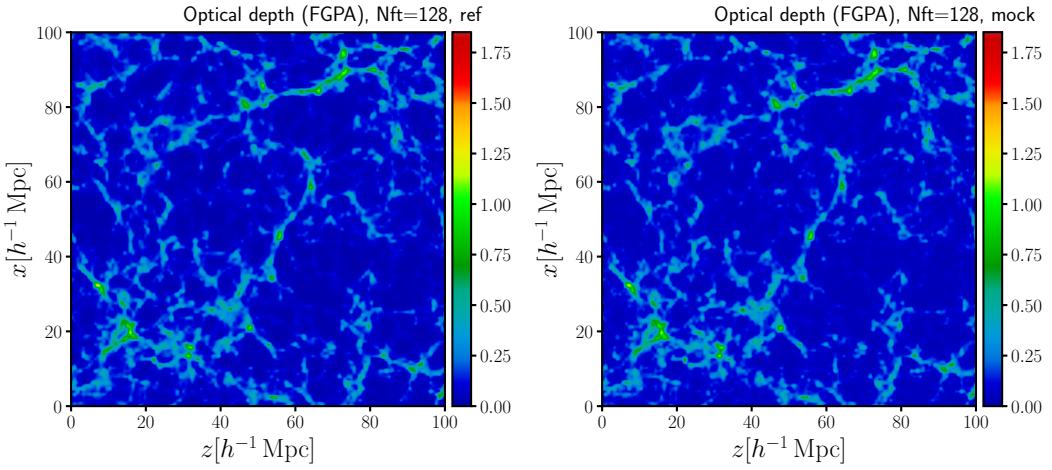

Figure 34: Slices of the reference simulation fields, extracted at the same position of the mesh. In the left panel, a slice of the reference τ_{FGPA} field, computed with the FGPA approximation; in the right panel a slice of the mock obtained with a simple local DM density bias model.

Mapping the FGPA optical depth

As final point of this chapter, we briefly present here the mapping of the optical depth τ_{FGPA} computed with the FGPA approximation. Because producing mocks with the FGPA approximation is not the real aim of this work, we limit to the study of the convergence of power spectrum, leaving the study of bispectrum out. The aim of this section is just to show that, if needed, `BAM` can handle the FGPA approximation in the fiducial convergence scenario. To this end, only a local DM density bias model is tested.

Figure 34 shows a comparison between the reference FGPA optical depth field (on the left) and the produced mock (on the right). The visual comparison shows a good agreement in the reproduction of the overall pattern of overdense structures. Figure 35 shows the comparison between the reference power spectrum (red solid line) and the mock power spectrum (green dashed line). The lower panel, which shows the ratio $P_{\mathrm{mock}}(k)/P_{\mathrm{ref}}(k)$ between mock and reference power spectrum, reveals that the mock power spectrum never deviates beyond 1% from the reference power spectrum. The average residuals R, computed up to $k = 0.7\,k_{\mathrm{nyq}}$, are $R \sim 0.40\%$, i.e. the convergence criterion is by far fulfilled.

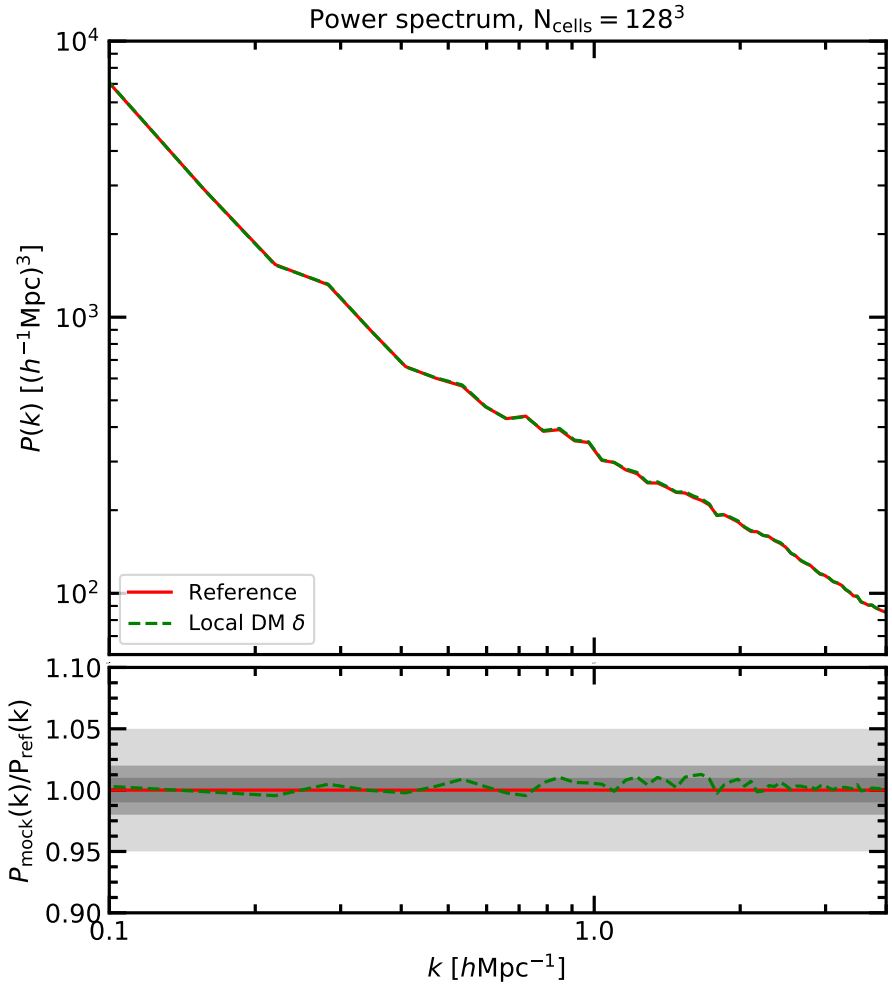

Figure 35: Comparison between reference (red solid line) and τ_{FGPA} mock power spectra, the latter obtained including in the bias the dependence only on the local DM overdensity (green dashed line). The upper panel shows the power spectra $P(k)$, while the lower panel shows the ratio between mock and reference power spectrum $P_{\rm mock}(k)/P_{\rm ref}(k)$ for the considered bias model. Gray shaded areas stand respectively, from the darker to the lighter, for $1\%, 2\%, 5\%$ residuals.

6 CONCLUSIONS AND OUTLOOKS

This master thesis presents new developments in mapping techniques between dark matter fields and various Large Scale Structure tracers, ranging from halos, over the ionised gas, the neutral hydrogen, the temperature of the intergalactic medium, to the Lyman-α forest.

In the beginning, we have explored new ways of characterising the complex cosmic web structure through the invariants of the gravitational tidal field tensor to reproduce the non-local anisotropic clustering components of the halo distribution.

Moreover, we have extended the work to be applied to hydrodynamical simulations to ultimately make accurate forward Lyman-α forest modelling.

In particular, have co-developed the `BAM` method to model the functional relationship between dark matter and baryonic quantities. This relation is complex, non-linear and non-local. But, since there is a non-vanishing correlation between the considered quantities, it is possible to define some bias expressed as a general tensor. When such relation has been identified, it can be extracted and calibrated numerically with `BAM`.

Let us be more specific about the work done in the different projects as follows.

In the first part of the work `BAM` has been applied to dark matter halos, in order to improve the mapping onto the dark matter field, already studied in previous papers. Concretely, the implementation of the invariants of the tidal field tensor $\mathcal{T}$ in the `BAM` code allowed to sample halo mock catalogs in a box of volume $V = 1500\,h^{-3}\mathrm{Mpc}^3$ and mesh resolution $l = 5h^{-1}\mathrm{Mpc}$, with unprecedent accuracy in the power spectrum at all scales and in the bispectrum on large scales, typically $k_1 = 0.1\,h\mathrm{Mpc}^{-1}$, $k_2 = 0.2\,h\mathrm{Mpc}^{-1}$. We dub I-web the ensemble of the invariants of $\mathcal{T}$. In particular, we find that the power spectrum improves with respect to previous results presented in Balaguera-Antolínez et al. (2019) and Pellejero-Ibañez et al. (2020) when I-web is included in the bias as a non-local dependence. The mock $P(k)$ hardly deviates more than 1% on large scales (Figures 10 and 11), which are more affected by the cosmic variance, and reduces significantly its scatter at small scales. On average, the mock power spectrum shows an unprecedented agreement with the reference power spectrum, within 0.5% up to $k = 0.72\,h\mathrm{Mpc}^{-1}$.

However, the most interesting results (shown in Figures 12 and 14) are obtained in the 3-point statistics, namely in the reduced bispectrum. We test the I-web model against other models inspired by Perturbation Theory. The results show that accounting for I-web in the bias produces mocks whose

bispectrum reproduces very accurately the bispectrum of the reference simulation at the investigated scales. The I-web model is in fact the unique which allows to obtain a mock bispectrum which is fully compatible within errors bars with the reference bispectrum. Moreover, such mock bispectrum is also the unique which cannot be rejected at all the investigated scales, according to a Bayesian analysis we performed. The net gain of information achieved with the I-web model is assessed by computing the significance of I-web with respect to a simple dark matter δ only bias model. Such significance turns out to be $\sim 4.8\,\sigma$: this is regarded as a clear detection of non-local bias. Moreover, the significance of I-web with respect to the commonly used T-web is $\sim 3.4\,\sigma$.

In parallel, we have shown that I-web represents an improved formulation of T-web, cosmic web classification based on the sign of the eigenvalues of the tidal field tensor. While T-web suffers the arbitrariness of the explicit choice of a threshold for the eigenvalues, I-web does not. Moreover, I-web can describe the information encoded in the cosmic web with an arbitrarily high number of bins, while in T-web this is restricted to just 4 bins (knots, filaments, sheets and voids).

As future expansions, because the bias is scale-dependent, we plan to go to smaller scales and study the bias at configurations such as $k_1 = 0.3\,h\mathrm{Mpc}^{-1}$, $k_2 = 0.6\,h\mathrm{Mpc}^{-1}$ in Fourier space. At such scales the I-web model may fail to reproduce the reference bispectrum with the accuracy achieved at the scales considered here. Therefore, we plan to investigate terms built from the velocity shear tensor and short-range non-local bias terms.

Another possible direction which can be taken is to investigate the connection between I-web and the scenario of galaxy formation and evolution. In fact, a number of correlations between cosmic web environments and galaxy properties are found. Therefore, one could think to deepen this topic by adopting the I-web perspective and look at how the invariants of the tidal field tensor and the other galaxy properties span the parameter space.

In the second part of the work we have applied the same mapping approach to the output of a zoom-in hydrodynamical simulation. This has been done in order to extract the bias relation between the dark matter field and its intergalactic baryonic tracers. The main results are the following.

The study of the cross-correlation reveals that dark matter highly correlates with ionized gas, $C(k) > 95\%$ at all scales. HI has a high cross-correlation with dark matter and gas on large scales, while on small scales the correlation slightly decreases. Moreover, the HI-gas cross-correlation on small scales is found to be higher than HI-DM cross-correlation. The temperature and Lyman-α photons optical depth correlate with dark matter on large scales, while on small scales the correlation is significantly reduced, becoming either vanishing or very low ($C(k) < 20\%$).

Differently from the case of dark matter halos, we have more than one tracer field available. Hence, we decide not to map each of those fields onto the dark matter field, but to adopt the perspective of extracting the mappings between baryonic fields too.

In particular, we start from the gas density and map it onto the dark matter field, due to their high cross-correlation. The analysis of the power spectrum does not reveal much on which are the crucial dependencies to be accounted for in the bias, because all the investigated bias models lead to a very well-converged mock power spectrum. However, the study of the bispectrum shows that short-range non-local bias terms, such as $\nabla^2\delta$, play an important role. Including $\nabla^2\delta$ in the bias allows in fact to reproduce the 3-point statistics with good accuracy. Instead, the dependence on the dark matter cosmic web is not as important as in the case of halos. Neither T-web nor I-web produce a significant improvement in the accuracy of bispectrum with respect to the result obtained with a local δ bias model. Bispectra studies, shown in Figure 19, yield quantitatively a significance of $\sim 2-3\,\sigma$ at all the investigated scales of the bispectrum obtained with $\nabla^2\delta$ with respect to the δ-only local case. This fact is interpreted, following the lines of chapter 4, as an evidence for short-range non-local bias. On the other hand, the significance of the bispectrum obtained with I-web with respect to the local δ model is within $1\,\sigma$. This confirms that long-range non-local dependencies do not produce significant improvements with respect to the local dependence.

Then, we map the HI number density onto the gas density. This perspective is quite interesting, because it allows to extract the mapping between two baryonic quantities, tracers of the dark matter field. We test separately local bias models based either on DM or on gas density and long-range non-local bias models consisting in DM I-web and gas I-web. The analysis of power spectrum suggests that the mapping on gas produces a better agreement than DM of the mock power spectrum with the reference power spectrum. Again, a strong evidence is found in the 3-point statistics. Indeed, locally the gas carries more information than DM on the spatial distribution of HI. Moreover, when the gas I-web is employed, the resulting mock bispectrum accurately reproduces the reference one. Reduced bispectra in Figure 22 reveal that the gas I-web model produces a HI bispectrum which is fully compatible with the reference bispectrum within errors bars at all scales. Moreover, a number of significancies are computed:

- $\sigma(\text{I-web gas, DM}\,\delta) \sim 2-3\,\sigma$: represents the total gain of information one achieves by mapping the HI onto the gas field and including local and non-local dependencies in the bias, with respect to the mapping based on local DM δ;

- σ(I-web gas, gas δ) $\sim 1.5\,\sigma$: represents the non-local gain of information achieved in mapping the HI onto the gas field when the gas I-web is accounted for in the bias;

- σ(I-web gas, I-web DM) $\lesssim 1\,\sigma$: represents the tiny gain of information one achieves non-locally when mapping the HI onto the gas field and not onto the DM field;

- σ(gas δ, DM δ) $\sim 1.5 - 2\,\sigma$: represents the local gain of information achieved by mapping the HI onto the gas instead of onto the DM field and taking into account only local dependencies.

As anticipated, the best result is obtained with the gas I-web bias model. However, also the dependence on DM I-web allows to produce quite accurate results. This is interpreted as follows.

According to the findings presented in Villaescusa-Navarro et al. (2018), namely most of the HI is located inside dark matter halos at $z < 5$, the DM-halos and DM-HI mappings are expected to be similar. Therefore, the DM I-web is expected to constrain the spatial correlations of HI on large scales, which are unfortunately out of scope given the volume of the simulation box. However, when going to the intermediate and small scales investigated here, the DM I-web fails to reproduce the 3-point statistics of HI in an accurate way. In fact, we find that the gas I-web achieves a better accuracy of the HI bispectrum than DM I-web. A possible explanation for this fact is that on such scales the gas I-web tracks the complex non-linear baryonic processes which involve gas and HI, but not DM.

The temperature requires adopting a different point of view. Indeed, when we try to map the temperature onto any of the other fields, the code does not converge in the fiducial scenario. The convergence of power spectrum gets stuck at average residuals $R \sim 50\%$. This is due to the fact that the temperature does not correlate very much with the DM, nor with gas or HI. This can be clearly seen in Figure 23.

To overcome this problem, we adopt the perspective of accounting for the photo-ionization equilibrium equation. Such equation expresses in this case the ratio between the ionized and the neutral atomic species as a function of temperature and the UV background. Therefore, we map the temperature onto the dark matter field, but explicitly include gas and HI in the bias as additional properties. In this way, when the code assigns a value of temperature it is able to take into account the relative abundance of ionized gas and HI, which partially constrains the temperature. By doing this, the results improve significantly. The mock temperature power spectrum deviates on average 5% from the reference temperature power spectrum (Figure 25). Moreover, we notice that the deviation is systematic. In fact, the mock $P(k)$ lies systematically below the reference power spectrum by a factor 5%, uniformly in the spherical shells in Fourier space and in a stable

way along iterations (Figure 27), once the maximum convergence is reached. By looking also at 3-point statistics, we suggest this nearly-constant bias may not have a physical origin, but it could be related to the numerical convergence. Therefore, we suggest to adopt a further convolution by a proper constant factor $\mathcal{K} > 1$, in order to shift uniformly upwards the mock power spectrum and make it fit the reference power spectrum. The final result of this operation is achieving $R \sim 1.01\%$, which is within the required convergence criterion. Even though this procedure has in principle a degree of arbitrariness, we regard the multiplying constant factor in Fourier space can be constrained case by case (see arguments in chapter 5), so that such arbitrariness is leveled out, at least partially. Figure 28 shows that the reference reduced bispectrum is very accurately reproduced by the mock bispectra.

The non-convolved and convolved bispectrum significancies with respect to the references are respectively $\lesssim 1\,\sigma$ and $\lesssim 0.5\,\sigma$. This means that, even in the case in which we do not post-process the mock with the convolution, the bias model accounting for photo-ionization equilibrium is physically well-motivated. Moreover, when applying the convolution, the accuracy of 3-point statistics in the mock improves. This means that artificially getting rid of the constant bias does not spoil the result obtained without the convolution. We stress that this mapping is very interesting, because for the first time a bias model accounting for dependencies on more than one field is presented.

The final step has been investigating the mapping between the optical depth for Lyman-α photons and the dark matter field in real space. In particular, thanks to the availability of the optical depth fields computed exactly by means of a proper integration of the HI number density along the line-of-sight, we have been able to compare the exact non-linear τ field with the optical depth field $\tau_{\rm FGPA}$, computed with the Fluctuating Gunn-Peterson Approximation (FGPA). We have showed that, at the investigated resolution, τ and $\tau_{\rm FGPA}$ have a high cross-correlation ($C(k) \sim 95\%$) on large scales, as expected, but a low-cross correlation on intermediate and small scales ($C(k) \sim 10\%$ at the Nyquist frequency). In terms of transmitted flux $F/F_c = \exp(-\tau)$, this translates into a different pattern of absorption lines. In particular, the FGPA approximation is shown to reproduce only the most prominent absorption features and to induce a shift of the lines with respect to the exact computation.

Subsequently, we have explored the possibility of mapping the exact τ directly onto the DM field. The first attempt is based just on a local DM density model and turns out to be unsuccessful. When one looks at the τ-DM cross-correlation, then it turns out that the correlation is high at large scales, but dramatically drops at small scales. This fact is interpreted as that the anisotropy introduced by the integration is not negligible. Therefore, we decide to try to modify the DM field and turn it into a fictious DM

optical depth $\tau_{\rm DM}$ by implementing a simplified and faster version of the integration along the line-of-sight. This results in a $\sim$ 20% increase of cross-correlation at small scales and in a $\sim$ 15% decrease of residuals of the mock power spectrum with respect to the reference power spectrum. On one hand, this attempt confirms the aforementioned argument about the anisotropy: it seems crucial to take into account the modifications that the integration along the line-of-sight introduces and which are not present in the FGPA approximation. On the other hand, even though $\tau_{\rm DM}$ is in principle unphysical, it paves the way to the strategy of mimicking the anisotropy effect with a simplified integration in the field onto which the exact τ is mapped. Indeed, one could think of repeating the same procedure e.g. with the HI field.

In the end, we have shown that, if needed, the `BAM` method can be successfully applied to map the $\tau_{\rm FGPA}$ optical depth, producing mock power spectra which meet the accuracy requirements.

The direct extension of this work is to refine the mapping between the Lyman-α forest and the dark matter field in real space (or eventually another baryon field, which in turn is mapped onto the dark matter, following a procedure similar to the one we have adopted for gas, HI and temperature) and then learn the mapping in redshift space. The outcome of such procedure will ultimately allow to generate a high number of accurate large-volume Lyman-α forest mocks, both in real and in redshift space, and analyze the cosmological information from the J-PAS and DESI surveys.

In the near future, we plan also to explore the suitability of `BAM` for producing mock catalogs of 21cm-line maps, in the view of the upcoming 21cm-line radio-surveys, to study the epoch of Reionization.

In the end, we have proven that the `BAM` method is able to produce not only very accurate dark matter halos mock catalogs, but also mock catalogs of the main properties of the intergalactic medium. Simultaneously, learning the mapping between dark matter and its tracers with `BAM` allows us to refine our comprehension of the bias on different cosmological scales. Moreover we have shown that the dark matter and baryonic cosmic webs carry a wealth of information on the spatial distribution of the tracers of the Large Scale Structure. Therefore, the cosmic web turns out to be a fundamental ingredient in the recipe of understanding our Universe.

APPENDIX A: MESH REPRESENTATION OF A FIELD

In this Appendix we provide some basic knowledge of how a field is represented on a mesh and how calculations are performed in the context of the grid discretization.

First, we introduce the so-called *Mass Assignment Schemes* (MAS hereafter), which consist in different window functions adopted for the interpolation of a set of particles on a mesh. Then, we present the discrete implementation of the FKP estimator of the power spectrum (see chapter 1).

1 MASS ASSIGNMENT SCHEMES

In general, two common computational ways to represent a field in a given volume is either as a set of discrete particles or as a set of values corresponding to elements of a mesh.

Let us consider a field, initially represented as a set of particles in a cubic box of side L and volume $V = L^3$. Such set of particles can be e.g. the output of a N-body simulation employing a tree-algorithm, or an initial distribution of particles (ICs) to be evolved in a simulation with the Particle Mesh technique. A set of n particles is in practice provided as a $n \times 3$-dimensional array, since each particles is defined by 3 spatial coordinates.

Suppose now we want to pass from the particle to the mesh representation. In this way, the field will no longer be treated as a set of particles, but as a grid of density values instead. This may be useful since in general many calculations (e.g. the computation of the power spectrum, which is very important in this work) are easier and faster in the mesh approach.

The way one passes from the particle to the mesh approach is by interpolating the set of particles on the cells of the grid with a proper kernel. In other words, one has to smooth the set of particles with a given window function and a smoothing length defined by the size of the cells. Even though this approach in general simplifies the calculations, it is a bit arbitrary, since the representation depends both on the adopted window function and on the smoothing length.

A number of window functions, commonly called *Mass Assignment Schemes* (MAS), have been presented and their performance and accuracy have been largely tested and debated in the literature.

We describe here the three most common MAS functions: the *Nearest Grid Point* (NGP), the *Cloud In Cell* (CIC) and the *Triangular Shaped Cloud*

(TSC). All the MAS are based on the idea of either assigning the mass budget of a particle entirely to a single cell or splitting and distributing it in a proper way among a group of cells.

The simplest MAS is the NGP, which assigns a particle entirely to the cell it belongs to. In real space it can be written as

$$\text{NGP:} \quad W(x_i) = \begin{cases} 1, & \text{if} \quad |x_i| < 0.5 \\ 0, & \text{else} \end{cases} \tag{61}$$

where x_i indicates the distance from the center of the i-th cell. NGP is commonly adopted to interpolate a set of particles which represent a field which is intrinsically discrete, like e.g. a catalog of DM halos or galaxies.

CIC and TSC are instead more commonly used to describe particles representing continuous density fields, such as dark matter or baryonic matter density fields. Differently from NGP, CIC and TSC split a particle in parts and re-distribute them among the cell in which such particle resides and the surrounding cells. This is done in order to make the resulting mesh field smoother and to level out, at least partially, the arbitrariness of the mesh representation.

CIC and TSC are computed in real space as follows:

$$\text{CIC:} \quad W(x_i) = \begin{cases} 1 - |x_i|, & \text{if} \quad |x_i| < 1 \\ 0, & \text{else} \end{cases}$$

$$\text{TSC:} \quad W(x_i) = \begin{cases} 0.75 - x_i^2, & \text{if} \quad |x_i| < 0.5 \\ (1.5 - |x_i|)^2/2, & \text{if} \quad 0.5 < |x_i| < 1.5 \\ 0, & \text{else} \end{cases}$$

In this work we adopt NGP or CIC (depending on the tracers) for the DM field, CIC for all the fields coming from the hydrodynamical simulation and NGP for DM halo fields.

The MAS can also be written as kernels in Fourier space, according to Hockney & Eastwood (1981):

$$W(k_i) = \left[\frac{\sin(\pi k_i)/2k_{\rm nyq}}{(\pi k_i)/2k_{\rm nyq}} \right]^p = \text{sinc} \left(\frac{\pi k_i}{2k_{\rm nyq}} \right)^p \quad , \tag{62}$$

where $k_{\rm nyq} = \pi\, N_{\rm cells}/L$ is the Nyquist frequency, k_i ($i = 1, 2, 3$) is the i-th component of wave-vector $\vec{k}$, and the exponent is $p = 1$ for NGP, $p = 2$ for CIC, $p = 3$ for TSC. Figure 36 shows in the left panel the three window functions in real space and in the right panel the square of the corresponding functions in Fourier space.

One non-trivial problem (among others, see e.g. Cui et al., 2008, and references therein) associated to the MAS in the computation of the power

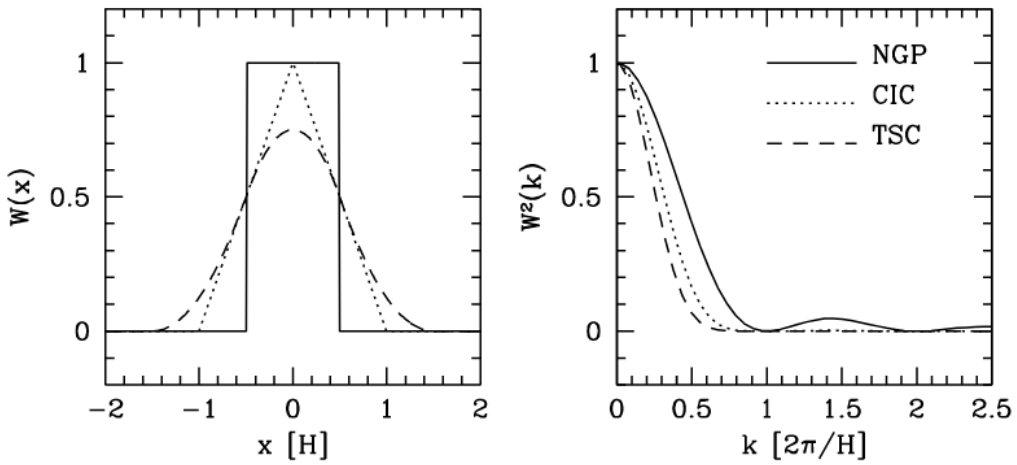

Figure 36: In the left panel, the three MAS window functions in real space: NGP (solid line), CIC (dotted line), TSC (dash-dotted line). In the right panel, the square of the corresponding functions in Fourier space. Plots from Cui et al. (2008).

spectrum is the so-called *aliasing effect*. Such effect arises due to the discreteness of the sampling on the grid and affects in different ways the MAS. In fact, the aliasing effect depends on the sampling and the MAS can be considered as a sampling on the grid of the product of the convolution of the field with the window. Therefore, using different MAS means sampling different objects on the grid.

Due to aliasing, power spectrum measurements relying on Fast Fourier Transforms (FFTs; discrete version of the Fourier Transforms) do not provide the true $P(k)$. In particular, from a mathematical point of view the $P(k)$ measurement will be reliable for $k \leqslant 0.5\,k_{\rm nyq}$, but we cannot trust the measurement for $k > 0.5\,k_{\rm nyq}$. Indeed, for $k > 0.5\,k_{\rm nyq}$, the aliasing results in a possible loss of power in the measurement (i.e. in an under-prediction) of $P(k)$ at high k. The $P(k)$ measurement of a NGP-interpolated field experiences a tiny loss of power, while CIC and TSC produces a severe under-estimation (TSC more than CIC) on small scales.

Therefore, in order to have reliable measurements of $P(k)$ one has to take into account this problem and eventually correct for aliasing. Even though a detailed treatment of aliasing corrections goes beyond the scope of this section, Jing (2005) showed that a proper deconvolution of the sampling effects of the window function, together with an iterative procedure, allows to recover the true power spectrum with 2% accuracy for $k \lesssim 0.7\,k_{\rm nyq}$.

2 ALGORITHM FOR POWER SPECTRUM MEASUREMENTS

In this section we present the discrete version of the FKP estimator (Feldman et al., 1994), already described in chapter 1. In practice, we just have to convert the continuous integral notation of chapter 1 into discrete sums.

Hence, we define the Fourier transform of the weighted galaxy (or halo, or any tracer) fluctuation field $F(\vec{k})$ as

$$F(\vec{k}) = \sum_g w(\vec{r}_g)\, e^{i\vec{k}\cdot\vec{r}_g} \; - \; \alpha \sum_s w(\vec{r}_s)\, e^{i\vec{k}\cdot\vec{r}_s} \tag{63}$$

where $w(\vec{r}_i)$ denotes the weighting functions and the subscript $i = \{g, s\}$ is related to the galaxy (or halo, or any tracer) catalog and to the uniform random catalog respectively.

The shot-noise correction is provided by

$$S(\vec{k}) = \alpha(1+\alpha) \sum_s w^2(\vec{r}_s)\, e^{i\vec{k}\cdot\vec{r}_s} \quad . \tag{64}$$

The 1-dimensional power spectrum, averaged on spherical shells in Fourier space, is finally estimated as

$$P(k) = \frac{1}{N_k} \sum_{k<|\vec{k}|<k+\delta k} |F(\vec{k})|^2 - S(0) \quad , \tag{65}$$

where the shot-noise correction $S(0)$ is in practice the reciprocal of the mean density (or mean number density, depending on the considered field) of the analyzed property inside the box. I.e., if we denote with $\bar{n} = N_{\rm halos}/L$ for instance the mean number density of halos, where $N_{\rm halos}$ is the total number of halos and L is the physical size of the box given in Mpc/h, then $S(0) = 1/\bar{n}$.

In our case, we set the thickness δk of each spherical shell in Fourier space equal to the fundamental mode $k_f = 2\pi/L$, i.e. $\delta k = k_f$.

APPENDIX B: ZOOM-IN PLOTS

In this Appendix we provide a full-page version of some plots already shown in chapter 5. This is done since in the printed version of this work such plots might result too small to be interpreted clearly.

Figures 37, 38 are related to Figure 19 (ionized gas bispectrum), Figures 39, 40 are related to Figure 22 (HI bispectrum), Figures 41, 42 are related to Figure 28 (temperature bispectrum).

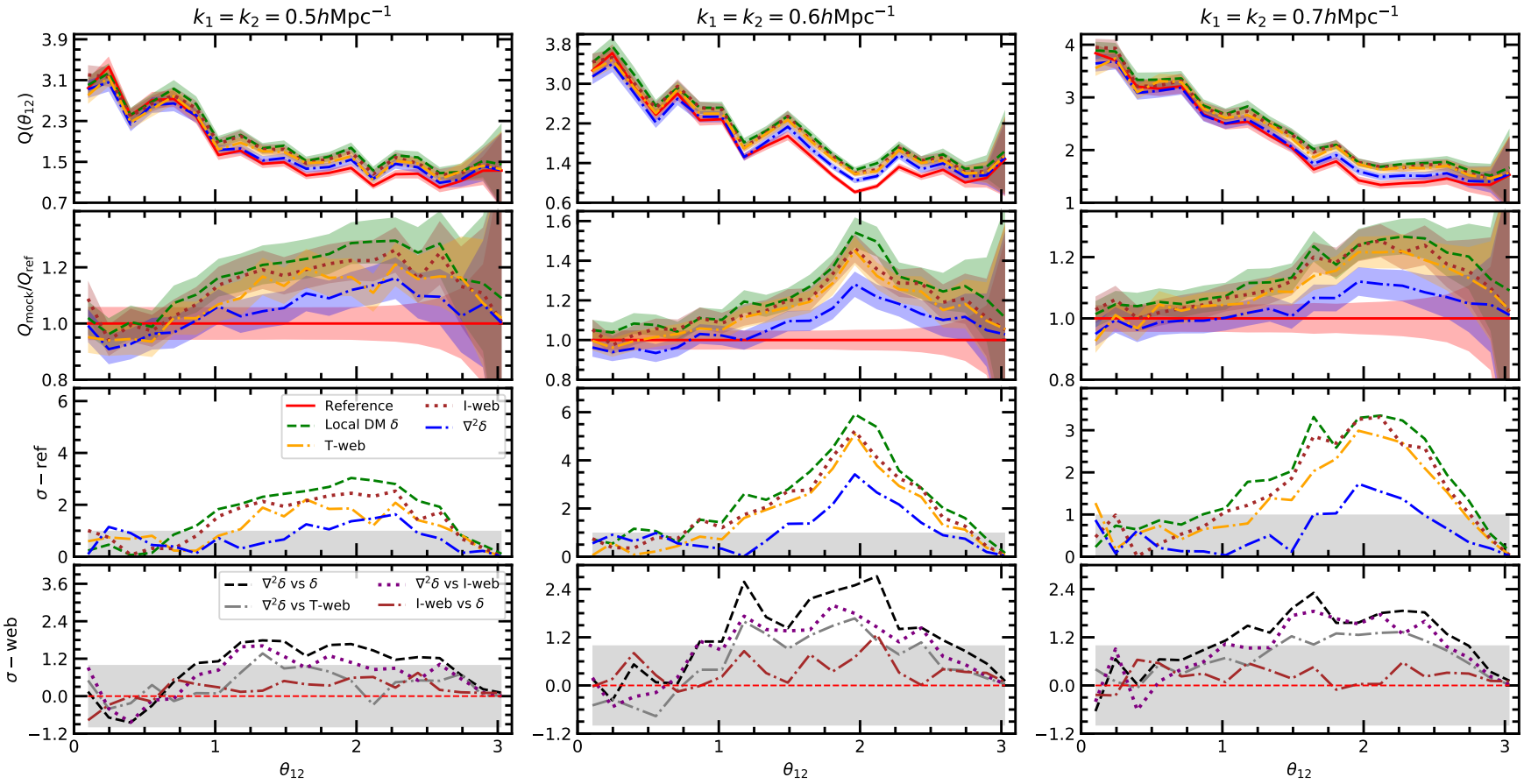

Figure 37: Reduced bispectra of reference field (red solid line) and gas mock catalogs, including in the bias respectively the dependence on local DM overdensity (green dashed line), the T-web (yellow dash-dotted line), the short-range non-local term $\nabla^2\delta$ (blue dash-dotted line) and I-web (brown dotted line). In the first row, reduced bispectra $Q(\theta_{12})$ and in the second row, ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$. Shaded areas represent error bars. In the third row and and fourth rows respectively the statistical significance $\sigma - \mathrm{ref}$ with respect to the reference and $\sigma - \mathrm{web}$ of a X bias model with respect to a Y one. Gray shaded areas stand for 1σ significance. The most accurate reproduction of the reference bispectrum is obtained including $\nabla^2\delta$ in the bias relation.

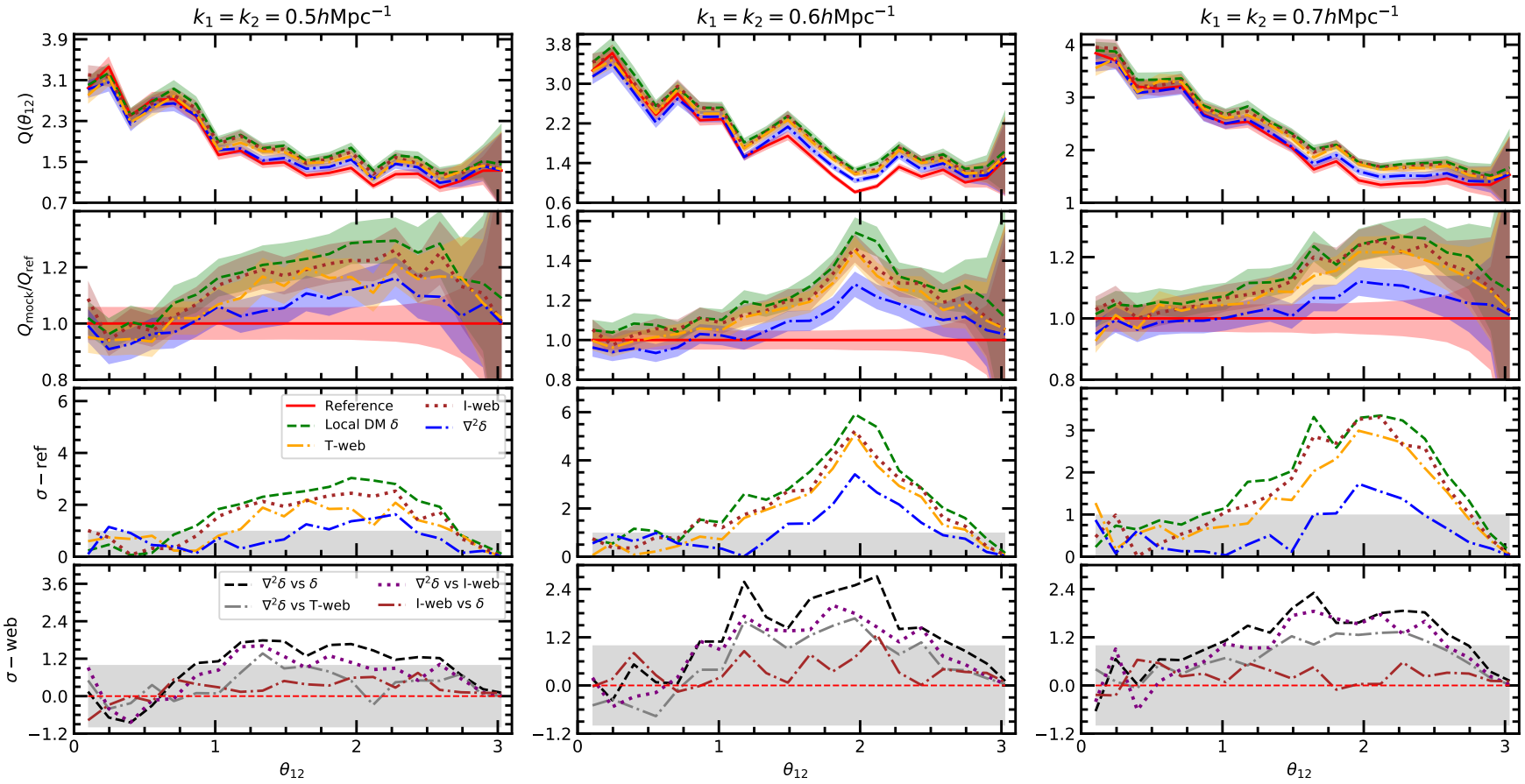

Figure 38: Reduced bispectra of reference field (red solid line) and gas mock catalogs, including in the bias respectively the dependence on local DM overdensity (green dashed line), the T-web (yellow dash-dotted line), the short-range non-local term $\nabla^2\delta$ (blue dash-dotted line) and I-web (brown dotted line). In the first row, reduced bispectra $Q(\theta_{12})$ and in the second row, ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$. Shaded areas represent error bars. In the third row and and fourth rows respectively the statistical significance $\sigma -$ ref with respect to the reference and $\sigma -$ web of a X bias model with respect to a Y one. Gray shaded areas stand for 1σ significance. The most accurate reproduction of the reference bispectrum is obtained including $\nabla^2\delta$ in the bias relation.

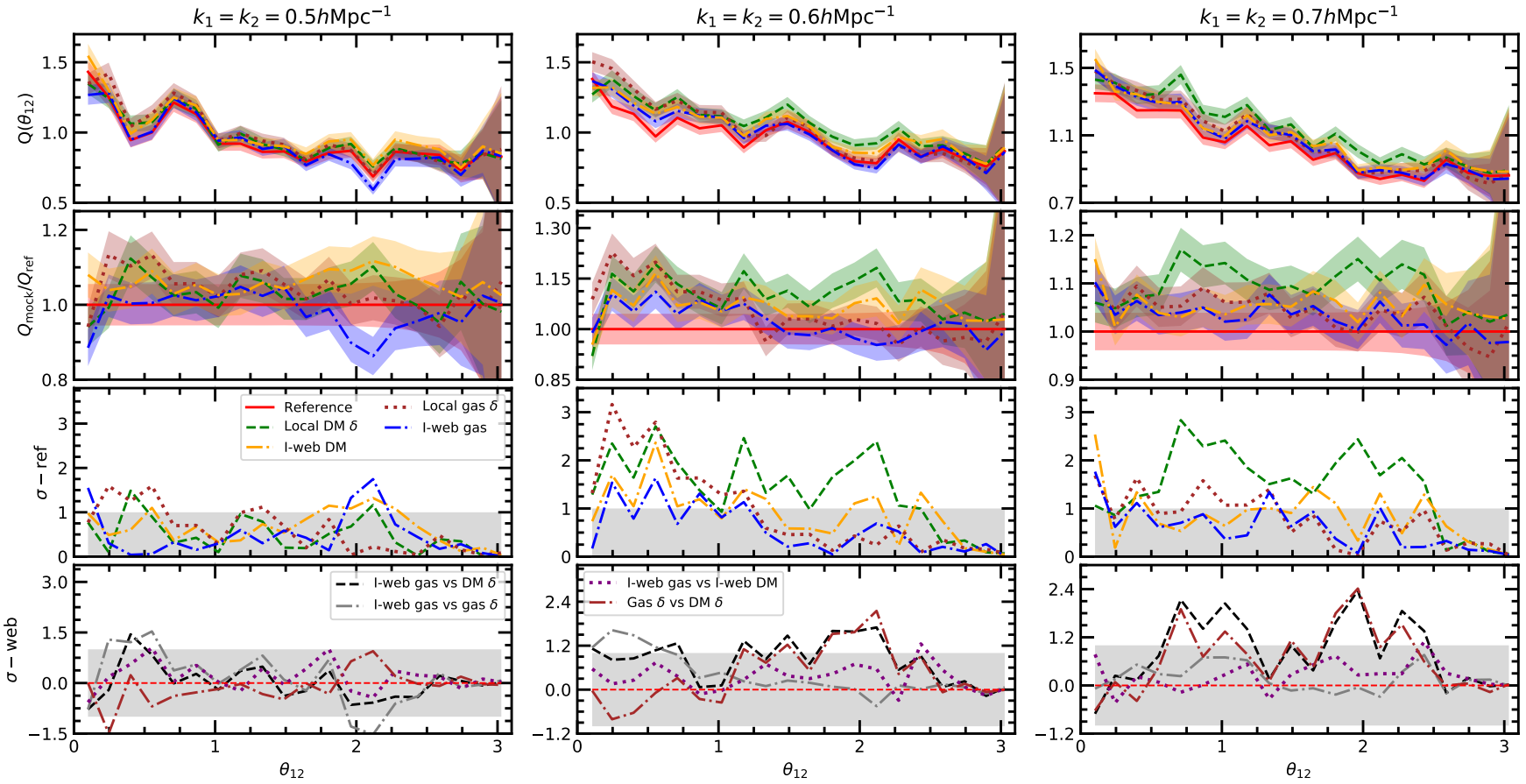

Figure 39: Reduced bispectra of reference field (red solid line) and HI mock catalogs, including in the bias respectively the dependence on local DM overdensity (green dashed line), DM I-web (yellow dashed line), local gas local overdensity (brown dotted line), gas I-web (blue dash-dotted line). In the first row, reduced bispectra $Q(\theta_{12})$ and in the second row, ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$. Shaded areas represent error bars. In the third row and and fourth rows respectively the statistical significance $\sigma - \mathrm{ref}$ with respect to the reference and $\sigma - \mathrm{web}$ of a X bias model with respect to a Y one. Gray shaded areas stand for 1σ significance. The most accurate reproduction of the reference bispectrum is obtained with gas I-web.

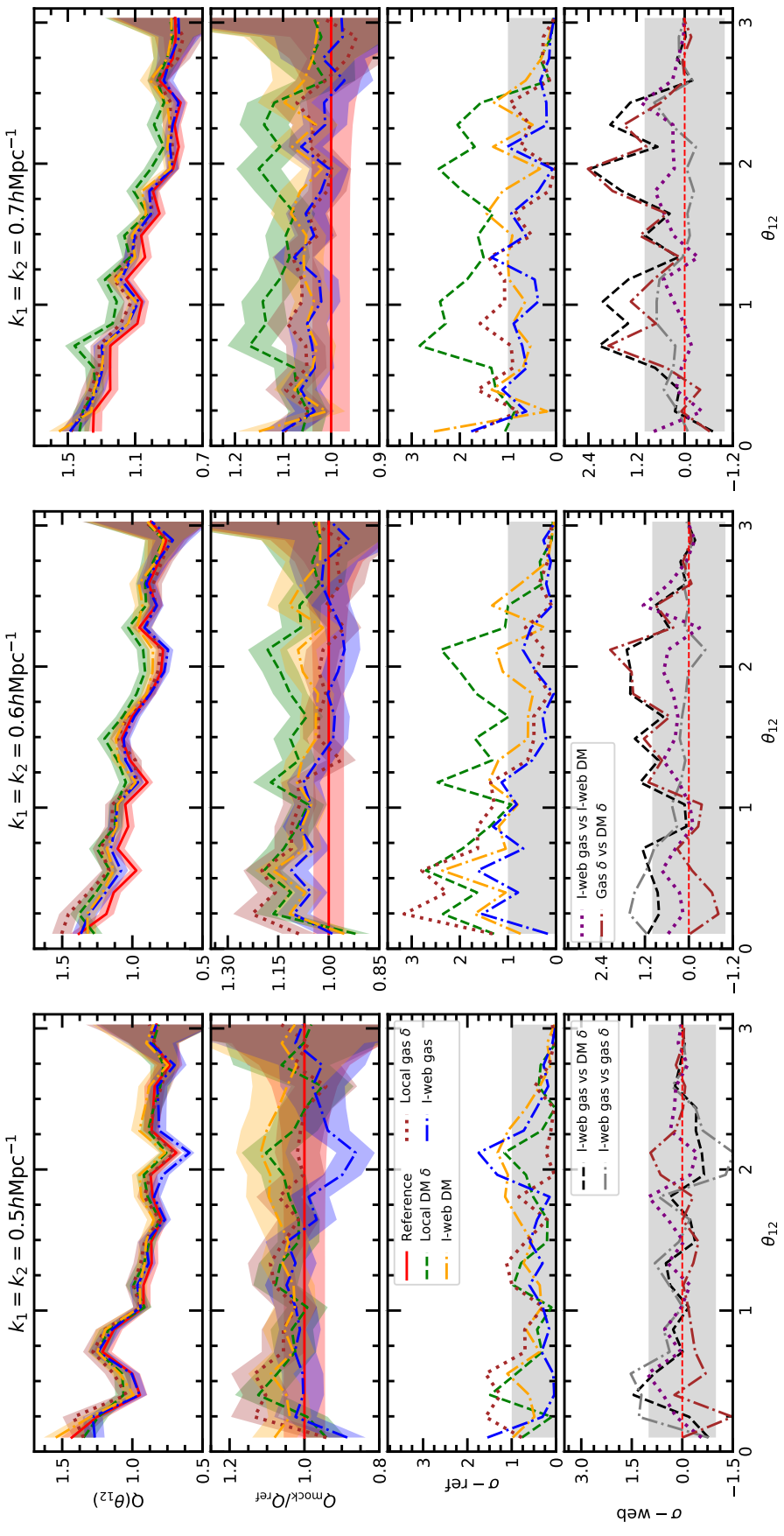

Figure 40: Reduced bispectra of reference field (red solid line) and HI mock catalogs, including in the bias respectively the dependence on local DM overdensity (green dashed line), DM I-web (yellow dashed line), local gas local overdensity (brown dotted line), gas I-web (blue dash-dotted line). In the first row, reduced bispectra $Q(\theta_{12})$ and in the second row, ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$. Shaded areas represent error bars. In the third row and and fourth rows respectively the statistical significance $\sigma - \mathrm{ref}$ with respect to the reference and $\sigma - \mathrm{web}$ of a X bias model with respect to a Y one. Gray shaded areas stand for 1σ significance. The most accurate reproduction of the reference bispectrum is obtained with gas I-web.

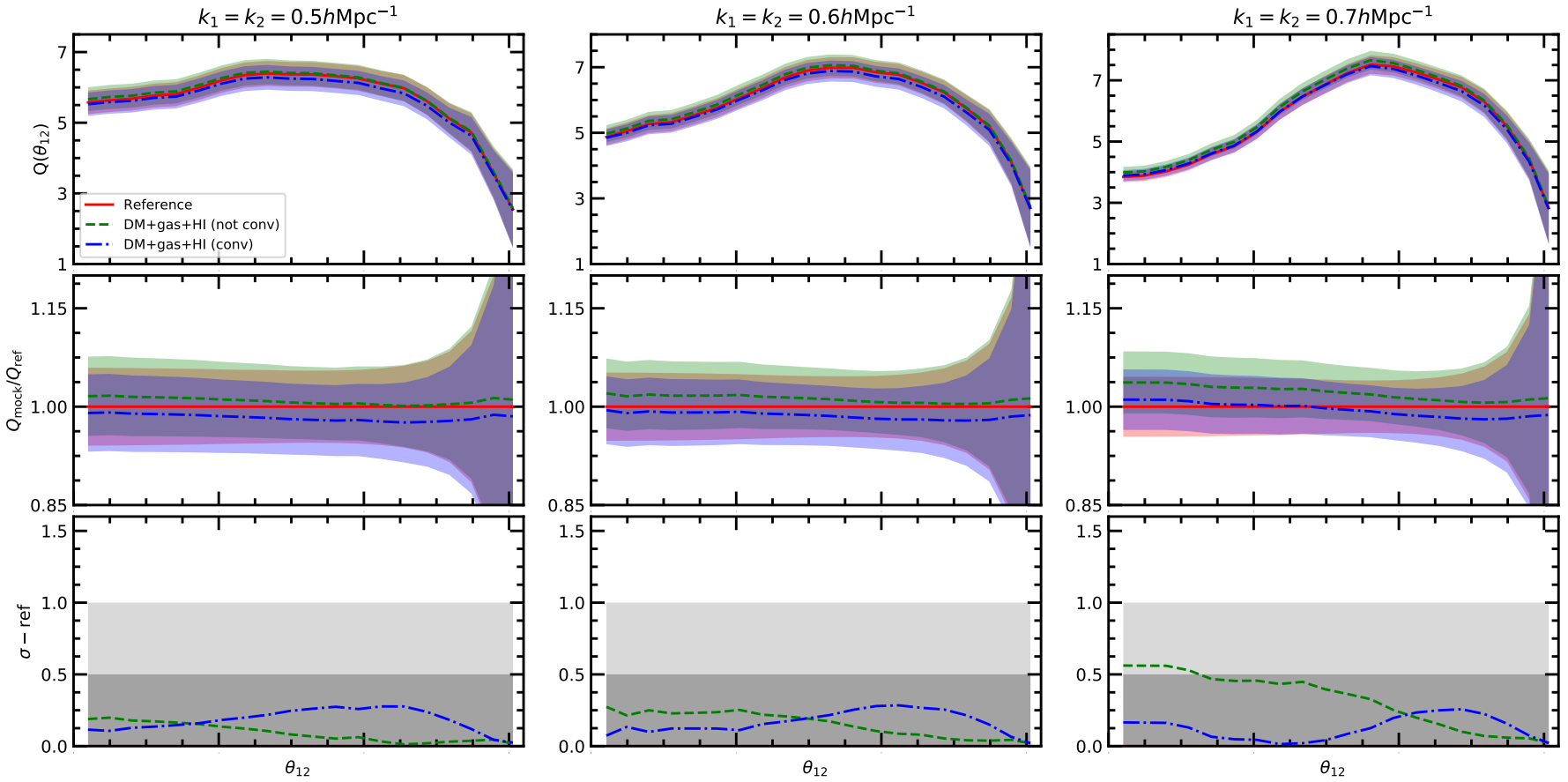

Figure 41: Reduced bispectra of reference field (red solid line) and temperature mock catalogs, including in the bias respectively the dependence on gas and HI overdensities (green dashed line) and the same + the convolution by $\mathcal{K} = \mathrm{const}$, $\forall k$ (blue dash-dotted line). In the first row, reduced bispectra $Q(\theta_{12})$ and in the second row, ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$. Shaded areas represent error bars. In the third row the statistical significance σ – ref with respect to the reference. Gray shaded areas stand for 1σ (lighter) and 0.5σ (darker) significance. The most accurate reproduction of the reference bispectrum is obtained by accounting for gas and HI overdensities in the bias and applying the further convolution.

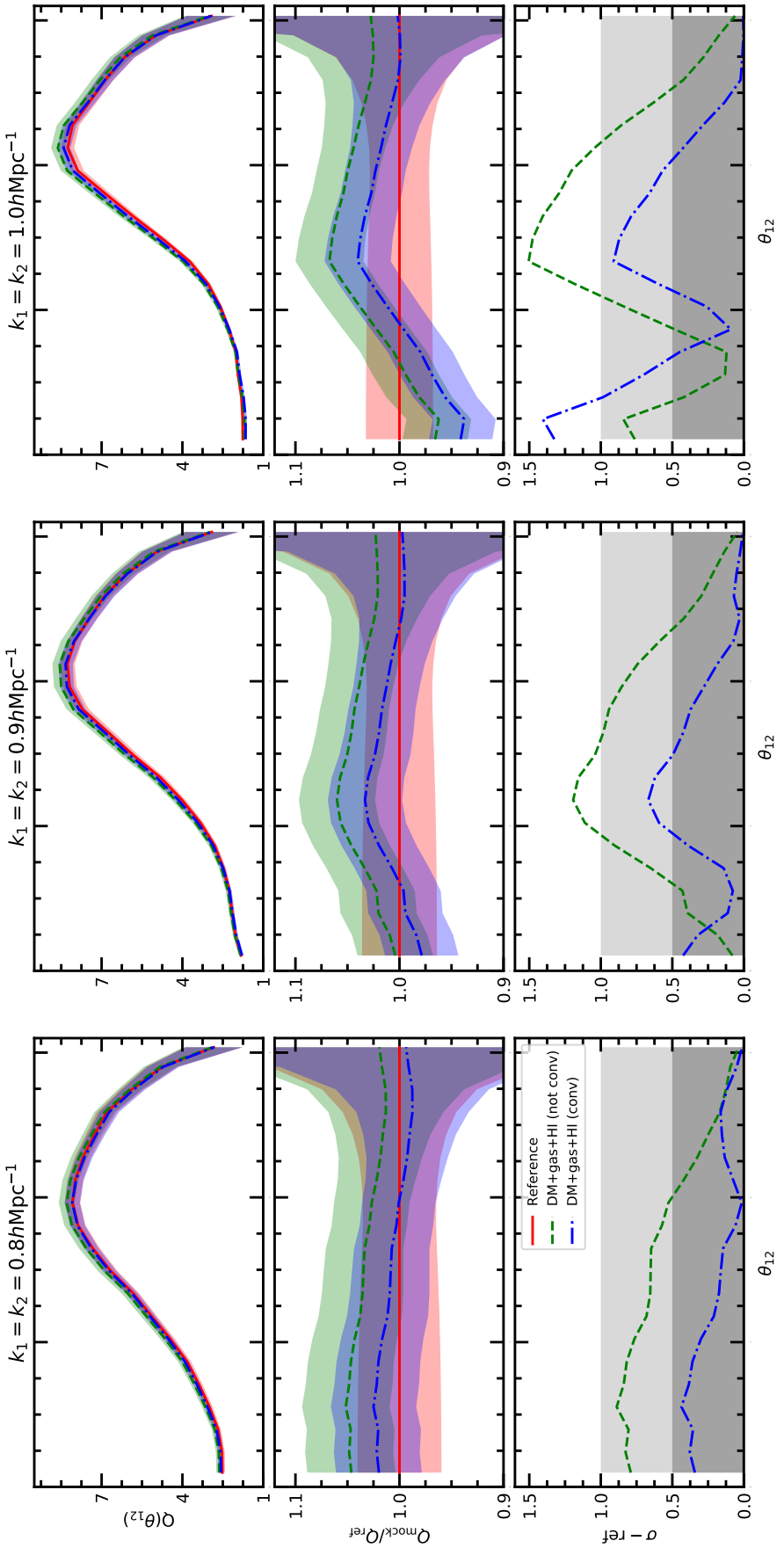

Figure 42: Reduced bispectra of reference field (red solid line) and temperature mock catalogs, including in the bias respectively the dependence on gas and HI overdensities (green dashed line) and the same + the convolution by $\mathcal{K} = \mathrm{const}$, $\forall k$ (blue dash-dotted line). In the first row, reduced bispectra $Q(\theta_{12})$ and in the second row, ratios $Q_{\mathrm{mock}}(\theta_{12})/Q_{\mathrm{ref}}(\theta_{12})$. Shaded areas represent error bars. In the third row the statistical significance $\sigma -$ ref with respect to the reference. Gray shaded areas stand for 1σ (lighter) and 0.5σ (darker) significance. The most accurate reproduction of the reference bispectrum is obtained by accounting for gas and HI overdensities in the bias and applying the further convolution.

BIBLIOGRAPHY

Abel, T., Hahn, O., & Kaehler, R. 2012, MNRAS, 427, 61. https://arxiv.org/abs/1111.3944

Amendola, L., Appleby, S., Avgoustidis, A., et al. 2018, Living Reviews in Relativity, 21

Angulo, R. E., Baugh, C. M., & Lacey, C. G. 2008, MNRAS, 387, 921. https://arxiv.org/abs/0712.2280

Angulo, R. E., Foreman, S., Schmittfull, M., & Senatore, L. 2015, Journal of Cosmology and Astroparticle Physics, 2015, 039–039

Aoyama, S., Hou, K.-C., Hirashita, H., Nagamine, K., & Shimizu, I. 2018, MNRAS, 478, 4905. https://arxiv.org/abs/1802.04027

Aragón-Calvo, M. A., Jones, B. J. T., van de Weygaert, R., & van der Hulst, J. M. 2007, Astronomy & Astrophysics, 474, 315–338

Avila, S., Murray, S., Knebe, A., et al. 2014, MNRAS, 450

Balaguera-Antolínez, A., Kitaura, F.-S., Pellejero-Ibáñez, M., Zhao, C., & Abel, T. 2018, MNRAS, 483, L58–L63

Balaguera-Antolínez, A., Kitaura, F.-S., Pellejero-Ibáñez, M., et al. 2019, MNRAS

Bardeen, J. M., Bond, J. R., Kaiser, N., & Szalay, A. S. 1986, ApJ, 304, 15

Blot, L., Crocce, M., Sefusatti, E., et al. 2019, MNRAS, 485, 2806. https://arxiv.org/abs/1806.09497

Bond, J. R., Kofman, L., & Pogosyan, D. 1996, Nature, 380, 603–606

Bond, J. R., & Myers, S. T. 1996, ApJ Supplement, 103, 1

Carilli, C., Gnedin, N., Furlanetto, S., & Owen, F. 2004, New Astronomy Reviews, 48, 1053–1061

Cautun, M., van de Weygaert, R., & Jones, B. J. T. 2012, Monthly Notices of the Royal Astronomical Society, 429, 1286–1308

Chabrier, G. 2003, Publications of the ASP, 115, 763. https://arxiv.org/abs/astro-ph/0304382

Chuang, C.-H., Kitaura, F.-S., Prada, F., Zhao, C., & Yepes, G. 2015, MNRAS, 446

Cieplak, A. M., & Slosar, A. 2016, Journal of Cosmology and Astroparticle Physics, 2016, 016–016

Colavincenzo, M., Sefusatti, E., Monaco, P., et al. 2019, MNRAS, 482, 4883, doi: 10.1093/mnras/sty2964

Contreras, S., Zehavi, I., Padilla, N., et al. 2019, MNRAS, 484, 1133. https://arxiv.org/abs/1808.02896

Crocce, M., & Scoccimarro, R. 2006, Physical Review D, 73

Cui, W., Liu, L., Yang, X., et al. 2008, The Astrophysical Journal, 687, 738–744

Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, The Astronomical Journal, 151, 44

Dekel, A., & Lahav, O. 1999, ApJ, 520, 24–34

Desjacques, V., Jeong, D., & Schmidt, F. 2018, Physics Reports, 733, 1–193

Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605. https://arxiv.org/abs/astro-ph/9709112

Elia, A., Kulkarni, S., Porciani, C., Pietroni, M., & Matarrese, S. 2011, MNRAS, 416, 1703. https://arxiv.org/abs/1012.4833

Farr, J., Font-Ribera, A., Bourboux, H. d. M. d., et al. 2020, Journal of Cosmology and Astroparticle Physics, 2020, 068–068

Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, The Astrophysical Journal, 426, 23

Feng, Y., Chu, M.-Y., Seljak, U., & McDonald, P. 2016, MNRAS, 463, 2273–2286

Forero-Romero, J. E., Hoffman, Y., Gottlöber, S., Klypin, A., & Yepes, G. 2009, Monthly Notices of the Royal Astronomical Society, 396, 1815–1824

Fry, J. N., Melott, A. L., & Shandarin, S. F. 1993, ApJ, 412, 504

Fry, J. N., & Peebles, P. J. E. 1978, ApJ, 221, 19

Gao, L., & White, S. D. M. 2007, MNRAS, 377, L5. https://arxiv.org/abs/astro-ph/0611921

GPy. since 2012, GPy: A Gaussian process framework in python, http://github.com/SheffieldML/GPy

Grieb, J. N., Sánchez, A. G., Salazar-Albornoz, S., & Dalla Vecchia, C. 2016, Monthly Notices of the Royal Astronomical Society, 457, 1577–1592

Gunn, J. E., & Peterson, B. A. 1965, ApJ, 142, 1633

Haardt, F., & Madau, P. 2012, ApJ, 746, 125. https://arxiv.org/abs/1105.2039

Hahn, O., & Abel, T. 2011, MNRAS, 415, 2101. https://arxiv.org/abs/1103.6031

Hahn, O., Abel, T., & Kaehler, R. 2013, MNRAS, 434, 1171. https://arxiv.org/abs/1210.6652

Hahn, O., Porciani, C., Carollo, C. M., & Dekel, A. 2007, MNRAS, 375, 489–499

Hockney, R. W., & Eastwood, J. W. 1981, Computer Simulation Using Particles

Hopkins, P. F., Narayanan, D., Murray, N., & Quataert, E. 2013, MNRAS, 433, 69. https://arxiv.org/abs/1209.0459

Hou, K.-C., Hirashita, H., Nagamine, K., Aoyama, S., & Shimizu, I. 2017, MNRAS, 469, 870. https://arxiv.org/abs/1704.01769

Hubble, E. 1929, Proceedings of the National Academy of Science, 15, 168

Hui, L., & Gnedin, N. Y. 1997, MNRAS, 292, 27, doi: 10.1093/mnras/292.1.27

Jeffreys, H. 1939, The Theory of Probability, Oxford Classic Texts in the Physical Sciences

Jing, Y. P. 2005, ApJ, 620, 559

Kaiser, N. 1987, MNRAS, 227, 1

Kitaura, F.-S., Angulo, R. E., Hoffman, Y., & Gottlöber, S. 2012, MNRAS, 425, 2422. https://arxiv.org/abs/1111.6629

Kitaura, F.-S., Balaguera-Antolínez, A., Sinigaglia, F., & Pellejero-Ibáñez, M. 2020, The cosmic web connection to the dark matter halo distribution through gravity. https://arxiv.org/abs/2005.11598

Kitaura, F.-S., Yepes, G., & Prada, F. 2014, MNRAS Letters, 439, L21–L25

Kooistra, R., Silva, M. B., & Zaroubi, S. 2017, Monthly Notices of the Royal Astronomical Society, 468, 857–869

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64

Levi, M., Bebek, C., Beers, T., et al. 2013, The DESI Experiment, a whitepaper for Snowmass 2013. https://arxiv.org/abs/1308.0847

Lewis, A., Challinor, A., & Lasenby, A. 2000, The Astrophysical Journal, 538, 473–476

Liddle, A. R. 1998, An introduction to modern cosmology

Lippich, M., Sánchez, A. G., Colavincenzo, M., et al. 2019, Monthly Notices of the Royal Astronomical Society, 482, 1786–1806

Matarrese, S., & Pietroni, M. 2007, JCAP, 2007, 026–026

Matarrese, S., Verde, L., & Heavens, A. F. 1997, MNRAS, 290, 651. https://arxiv.org/abs/astro-ph/9706059

McDonald, P., & Eisenstein, D. J. 2007, Physical Review D, 76

McDonald, P., & Roy, A. 2009, Journal of Cosmology and Astroparticle Physics, 2009, 020–020, doi: 10.1088/1475-7516/2009/08/020

Monaco, P., Theuns, T., & Taffoni, G. 2002, MNRAS, 331, 587–608

Morris, J. P. 1996, Publications of the Astron. Soc. of Australia, 13, 97

Nagamine, K., Shimizu, I., Fujita, K., et al. 2020, Probing Feedback via IGM tomography and Ly-alpha forest with Subaru PFS, TMT/ELT & JWST. https://arxiv.org/abs/2007.14253

Obuljen, A., Dalal, N., & Percival, W. J. 2019, Journal of Cosmology and Astroparticle Physics, 2019, 020

Peebles, P. J. E. 1980, The large-scale structure of the universe

Pellejero-Ibañez, M., Balaguera-Antolínez, A., Kitaura, F.-S., et al. 2020, MNRAS, 493, 586–593

Perlmutter, S., Aldering, G., Goldhaber, G., et al. 1999, ApJ, 517, 565–586

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A13. https://arxiv.org/abs/1502.01589

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2018, Planck 2018 results. VI. Cosmological parameters. https://arxiv.org/abs/1807.06209

Popping, A., Meyer, M., Staveley-Smith, L., et al. 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14), 132. https://arxiv.org/abs/1501.01077

Porqueres, N., Jasche, J., Lavaux, G., & Enßlin, T. 2019, Astronomy & Astrophysics, 630, A151

Press, W. H., & Schechter, P. 1974, ApJ, 187, 425

Rasmussen, C. E., & Williams, C. K. I. 2005, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning) (The MIT Press)

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, The Astronomical Journal, 116, 1009–1038

Saitoh, T. R. 2016, CELib: Software library for simulations of chemical evolution. http://ascl.net/1612.016

—. 2017, Astronomical Journal, 153, 85. https://arxiv.org/abs/1612.02260

Saitoh, T. R., & Makino, J. 2013, ApJ, 768, 44. https://arxiv.org/abs/1202.4277

Schmittfull, M., Simonović, M., Assassi, V., & Zaldarriaga, M. 2019, Physical Review D, 100

Scoccimarro, R. 2000, The Astrophysical Journal, 544, 597–615

Scoccimarro, R. 2015, Physical Review D, 92, 083532. https://arxiv.org/abs/1506.02729

Scoccimarro, R., Colombi, S., Fry, J. N., et al. 1998, The Astrophysical Journal, 496, 586–604

Scoccimarro, R., & Sheth, R. K. 2002, MNRAS, 329, 629–640

Sefusatti, E., Crocce, M., Pueblas, S., & Scoccimarro, R. 2006, Physical Review D, 74

Sefusatti, E., Crocce, M., Scoccimarro, R., & Couchman, H. M. P. 2016, MNRAS, 460, 3624. https://arxiv.org/abs/1512.07295

Seljak, U. 2012, JCAP, 2012, 004–004

Sheth, R. K., Mo, H. J., & Tormen, G. 2001, Monthly Notices of the Royal Astronomical Society, 323, 1–12

Shimizu, I., Todoroki, K., Yajima, H., & Nagamine, K. 2019, MNRAS, 484, 2632. https://arxiv.org/abs/1901.03815

Smith, B. D., Bryan, G. L., Glover, S. C. O., et al. 2017, MNRAS, 466, 2217. https://arxiv.org/abs/1610.09591

Sousbie, T., Courtois, H., Bryan, G., & Devriendt, J. 2008, ApJ, 678, 569

Springel, V. 2005, MNRAS, 364, 1105. https://arxiv.org/abs/astro-ph/0505010

Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726. https://arxiv.org/abs/astro-ph/0012055

Tasitsiomi, A. 2006, The Astrophysical Journal, 648, 762

Tassev, S., Zaldarriaga, M., & Eisenstein, D. J. 2013, JCAP, 2013, 036–036

The Dark Energy Survey Collaboration. 2005, The Dark Energy Survey. https://arxiv.org/abs/astro-ph/0510346

Tinker, J. L., Robertson, B. E., Kravtsov, A. V., et al. 2010, ApJ, 724, 878. https://arxiv.org/abs/1001.3162

Villaescusa-Navarro, F., Genel, S., Castorina, E., et al. 2018, The Astrophysical Journal, 866, 135

Weinberg, D. H., Hernsquit, L., Katz, N., Croft, R., & Miralda-Escudé, J. 1997, in Structure and Evolution of the Intergalactic Medium from QSO Absorption Line System, ed. P. Petitjean & S. Charlot, 133. https://arxiv.org/abs/astro-ph/9709303

Weltman, A., Bull, P., Camera, S., et al. 2020, Publications of the Astronomical Society of Australia, 37

Werner, K. F., & Porciani, C. 2019, Monthly Notices of the Royal Astronomical Society, 492, 1614–1633

White, M., Tinker, J. L., & McBride, C. K. 2014, MNRAS, 437, 2594. https://arxiv.org/abs/1309.5532

White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341

Zentner, A. R. 2007, International Journal of Modern Physics D, 16, 763. https://arxiv.org/abs/astro-ph/0611454

CONTRIBUTIONS OF THE AUTHOR

As a clarification, the contributions of the author are here listed.

In the first part of the work, related to the assessment of non-local halo bias models (chapter 4), I performed all the `BAM` runs and the calculations of summary statistics, i.e. of power spectrum and bispectrum, including the assessment of errors and significance. Moreover, I made all the plots, contributed with the improved treatment of variables and helped to de-bug the code. The Bayesian analysis of errors has been performed by Dr. Marcos Pellejero-Ibañez. The analytical calculations which link the T-web and I-web formulation of the cosmic web has been performed by Prof. Francisco-Shu Kitaura, who also had the inspiration of investigating the connection between phenomenological and PT approaches. The `BAM` code has been mainly developed by Dr. Andrés Balaguera-Antolińez. The theoretical formalization, discussion and understanding of the bias models is the result of a fruitful team-work, in which I took part. The results of such team-work led to the submission of Kitaura et al. (2020) in May 2020.

In the second part of the work, regarding the mapping of a hydrodynamical simulation on the dark matter field, I again performed all the `BAM` runs and calculations of summary statistics and made all the plots. In particular, I contributed with the idea of exploiting the mapping between baryonic fields (map of HI on gas) and to extract the information regarding the baryonic components of the cosmic web (gas I-web). Moreover, I implemented the bias dependencies on gas and HI, so as to constrain the spatial temperature field with the photo-ionization equilibrium. Then, I computed the FGPA approximation and compared it with the exact computation of the optical depth. Finally, I contributed with the idea of adding an anisotropy in the dark matter field, in order to achieve a more accurate bias mapping of the optical depth, and tested such idea in the code. The extension of `BAM` to continuous field has been implemented by Manuel Sańchez Benavente, who also performed an analysis of the convergence of power spectrum of gas, HI and temperature in his Master Thesis.

As complementary part of this work, I coded a `Python` version of the I-web computation, to be addressed to the analysis of observational galaxy catalogs, in order to investigate the correlations between galaxy properties and the cosmic environment in an I-web perspective.

ACKNOWLEDGMENTS

This work marks the end of a 5-year long adventure, which I spent partly in Padova, my town, and partly in Tenerife, in the Canary Islands. There have been several difficult moments along the way, but at the end of the day successes and satisfactions largely overwhelmed disappointments, so that I can say I had had a very great time. And, above all, I enjoyed very much my Master Thesis, which is the ultimate goal of the study and the work of the last 2 years.

Many people have contributed to reach the finish line. I would like to take the opportunity to acknowledge them all now, in the language which most suits the context.

The first heartfelt dutiful acknowledgment goes to Prof. Francisco-Shu Kitaura and Dr. Andrés Balaguera-Antolińez. Paco and Andrés welcomed me with open arms in their research group at IAC and made me feel like at home. Both of them have dedicated to me much of their time, to teach me new things, answer my questions, clarify my doubts and help me with the code and the writing.

I would like to thank personally Paco for supporting me very strongly throughout this experience, involving me unreservedly in many exciting projects, and teaching me a very important way of thinking: no matter how steep the road might be, never give up, keep on fighting for your ideas if you believe in them.

I would like to thank personally also Andrés for teaching me innumerable useful and interesting things, of Astrophysics and coding, and for dedicating much time to help me with the issues I encountered, with great patience, always smiling (I must admit I was very bad at C++ when I begun).

Lastly, I would like to thank both of them for their warmth and friendliness, which contributed to create an optimal working environment. *Muchas gracias!*

I take the opportunity to thank also other people who contributed to this Master Thesis. I would like to thank Prof. Kentaro Nagamine and his research group, for kindly providing their hydrodynamical simulation. The part of this work regarding the bias of baryonic fields would not have been possible otherwise. I would like to thank also Manuel, for providing me the last version of the code he had been using and for useful discussions throughout the work.

I thank also Prof. Michele Liguori, for supporting me in doing my thesis abroad and for his availability to be my internal supervisor.

Dopo i ringraziamenti scientifici, è doveroso fare una serie di ringraziamenti personali.

Il primo ringraziamento va ai miei genitori, Myriam e Stefano, il mio porto sicuro nei momenti difficili, che mi hanno sostenuto in modo incondizionato

durante tutto il percorso e senza i quali tutto questo in fondo non sarebbe stato possibile. Vi dedico questo lavoro, che è tanto mio quanto vostro.

Un ringraziamento di cuore va ai miei fratelli, Edoardo e Margherita, fonti delle più belle e sincere risate e punti di riferimento fondamentali nella mia vita.

Un ringraziamento speciale va ovviamente alla mia Gaia, sempre immancabilmente al mio fianco in ogni momento, compagna di mille splendide avventure in giro per il mondo (anche se continui a sostenere che la Cosmologia non esiste!).

Molti sono gli amici che vorrei ringraziare. Siete tanti e siete tutti importanti, menzionerò solo alcuni, ma sappiate che gli altri hanno un posto altrettanto importante nel mio cuore:

gli amici del Luvro, per la loro amicizia ormai più che quindicennale: sarebbe impossibile elencare tutti i motivi per cui vi potrei ringraziare. Tuttavia, fatemi menzionare la mitica partita di Scarabeo che ha coniato il nome del gruppo, finita (ovviamente) in litigio, grazie alla quale abbiamo scoperto che il pesce fragolino (anche detto Luvaro, o Luvro) sembra essere delizioso;

gli amici del quartetto sospettosi, Giovanni, Gregorio e Nicola, per le chiacchierate ad Albi e i gelati da nonno Piero;

los amigos del Erasmus, Erica, Sophie, Julia, Riccardo e Angela, para los viajes en las islas Canarias y las adventuras tinerfeñas;

i tanti amici astronomi, che tra lezioni in Rosino, pranzi in Murialdo, caffè di metà mattina, chiacchiere e partite di briscola in pausa pranzo (una tra le altre, la briscola in control room del telescopio di Asiago), hanno reso questo percorso indimenticabile;

los astrofisićos de la ULL, para la ayuda con Python en el CCA, los preciosos senderismos en Anaga que empiezan en las montañas y acaban en la playa, las hamburguesas con batidos de fruta y las charlas en el café 7;

tutti gli amici musicisti, con i quali fare della bella musica è sempre una cura per l'anima;

mis compañeros de piso en cuarentena, Nico y Francesco, para los partidos de Carcassonne y de poker, apostando platos que fregar: creo que aún alguien me debe unos platos;

tutti quelli che, in varia misura, hanno contributo a raggiungere questo traguardo.

Già che ci siamo, ringrazio anche le isole Canarie, Tenerife e papà Teide in primis, per avermi accolto come una seconda casa e per essere un posto così splendido.

Infine, un ringraziamento un po' particolare va al nostro Universo, per essere così misteriosamente affascinante.

Grazie a tutti, *hasta la proxima!*