

# Going beyond the galaxy power spectrum: An analysis of BOSS data with wavelet scattering transforms

Georgios Valogiannis<sup>\*</sup> and Cora Dvorkin<sup>†</sup>

*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*

 (Received 6 May 2022; accepted 26 October 2022; published 14 November 2022)

We perform the first application of the wavelet scattering transform (WST) to actual galaxy observations, through a WST analysis of the BOSS DR12 CMASS dataset. We included the effects of redshift-space anisotropy, nontrivial survey geometry, systematic weights, and the Alcock-Paczynski distortion effect, following the commonly adopted steps for the power spectrum analysis. In order to capture the cosmological dependence of the WST, we use galaxy mocks obtained from the state-of-the-art ABACUSSUMMIT simulations, tuned to match the anisotropic correlation function of the BOSS CMASS sample in the redshift range  $0.46 < z < 0.60$ . Using our model for the WST coefficients, as well as for the first 2 multipoles of the galaxy power spectrum, that we use as reference, we perform a likelihood analysis of the CMASS data. We obtain the posterior probability distributions of four cosmological parameters,  $\{\omega_b, \omega_c, n_s, \sigma_8\}$ , as well as the Hubble constant, derived from a fixed value of the angular size of the sound horizon at last scattering measured by the *Planck* satellite, all of which are marginalized over the seven nuisance parameters of the halo occupation distribution model. The WST is found to deliver a substantial improvement in the values of the predicted  $1\sigma$  errors compared to the regular power spectrum, which are tighter by a factor of 3–5 in the case of flat and uninformative priors and by a factor of 3–8, when a big bang nucleosynthesis prior is applied on the value of  $\omega_b$ . Our results are investigative and subject to certain approximations, which we discuss in the text.

DOI: [10.1103/PhysRevD.106.103509](https://doi.org/10.1103/PhysRevD.106.103509)

## I. INTRODUCTION

The process of gravitational instability, modulated by the expansion of the universe and the set of fundamental interactions between its basic constituents, has led to the emergence of the large-scale structure (LSS) of the universe out of the primordial cosmic density field. The observed 3-dimensional (3D) distribution of matter is, as a result, a powerful probe of fundamental physics, which can reveal a wealth of information about the nature of the late-time accelerated expansion of the universe [1], the law of gravity at large scales [2–4], the nature of dark matter [5], the properties of massive neutrinos [6,7] and the physics of the early universe [8]. Aiming to tap into this valuable resource, a wide array of current and future cosmological surveys, such as the Dark Energy Spectroscopic Instrument (DESI) [9], the *Vera C. Rubin* Observatory Legacy Survey of Space and Time (LSST) [10,11], Euclid [12], and the *Nancy Grace Roman* Space Telescope [13], among others, will trace the galaxies in the cosmic web with unprecedented levels of accuracy, potentially allowing us to explore and test the vast landscape of cosmological scenarios that propose to tackle these unresolved questions.

On the theory front, the expected influx of data needs to be complemented by an associated theoretical framework to quantify and extract all possible cosmological information encoded in the LSS of the universe, a step that is underway. One of the most efficient ways to extract information out of an observed dataset is through the evaluation of the 2-point correlation function or its Fourier-space counterpart, the power spectrum. Combined with theoretical modeling and/or state-of-the-art simulations to capture the cosmological dependence of the target statistic and its covariance matrix, the values of the cosmological parameters that constitute a given scenario can be determined, up to a certain degree of accuracy. Despite serving as a useful first line of attack in problems of cosmological parameter inference, the power spectrum analysis is known to be incomplete, because it fails to capture a significant part of the information content in the LSS: the non-Gaussian part of the distribution sourced by the process of gravitational instability that drives structure formation [14]. In order to fully exploit the additional gains associated with tapping into the nonlinear regime of the LSS, one thus needs to evaluate higher-order statistics beyond the 2-point function. Despite significant theoretical progress made in this direction in the past decade [15–21], including higher-order moments as part of a standard parameter inference scheme quickly becomes intractable, both from a computational standpoint, because of the sharply rising

<sup>\*</sup> [gvalogiannis@g.harvard.edu](mailto:gvalogiannis@g.harvard.edu)

<sup>†</sup> [cdvorkin@g.harvard.edu](mailto:cdvorkin@g.harvard.edu)

dimensionality of the resulting data vector, but also on the theoretical modeling front. In addition, in more challenging cases, such as that of a probability distribution with a heavy tail, even a complete description of all moments would fail to capture all available information, while at the same time amplifying outliers by raising the density field to very high powers [14].

Furthermore, realistic probes of the LSS, such as the 3D pattern of galaxies observed by spectroscopic surveys, do not generally perfectly trace the underlying matter distribution, but are rather *biased* tracers of it [22], due to the complicated physics of structure formation. In addition, when galaxies are identified through spectroscopic means, their peculiar velocities around the Hubble flow lead to a perceived anisotropy of the observed clustering pattern, the redshift-space distortions (RSD) [23]. Combined with other challenges associated with systematic errors or galaxies observed in a nontrivial survey geometry [24], an additional layer of complexity is added that makes the modeling and interpretation efforts more difficult.

Aiming to overcome the former challenge, an active area of research is focused on developing novel estimators beyond the 2-point function, attempting to access higher-order information without having to explicitly evaluate the full correlation hierarchy. This broad spectrum of approaches consists of, but is not limited to, attempts to Gaussianize the density distribution by suppressing the contribution of LSS from regions with high overdensity [14,25–29], estimators that harness the information from regions that have not undergone nonlinear gravitational collapse, the cosmic voids [30–36], proxy lower-order statistics [37–39], and a variety of other statistics beyond the power spectrum, such as Minkowski functionals [40–43], the  $k$ -nearest neighbor cumulative distribution functions [44], the minimum spanning tree [45] or 1-point statistics [46]. More recently, convolutional neural networks (CNNs) [47] have emerged as a completely novel approach that promises to reliably identify features of complex datasets, including a potentially full extraction of their non-Gaussian information content. Despite very promising results on cosmological applications [48–51], the extent to which their outcomes can be interpreted in order to allow reliable applications on real galaxy data is still an open question.

In our recent work [52], we investigated the prospect of bridging the gap between the use of traditional estimators and CNNs in modern cosmological analyses of the LSS, using the wavelet scattering transform (WST) estimator [47,53]. Originally proposed in the context of signal processing in computer vision, the WST subjects an input physical field to a series of successive nonlinear operations (wavelet convolution and modulus), in a network architecture that resembles the one of a CNN with fixed kernels [47,53–56]. The outcome of this process is a compact set of a few coefficients that can serve as a basis that reflects the clustering properties of the input field beyond the 2-point

function, without raising the field to very high powers, a common shortcoming of traditional estimators [14], while at the same time retaining its interpretability, unlike in CNNs [54–56]. In light of the great promise held in the use of this estimator, the WST has been recently applied in the fields of astrophysics [57–59], cosmology [52,60–64] and molecular chemistry [65,66] (a review can be found in Ref. [67]).

Through the first WST application to 3D matter density fields, simulated by the QUIJOTE simulations [68], we showed in Ref. [52] how the WST can deliver a very large improvement in the extracted errors on cosmological parameters. Motivated by these promising results, in this work we carry out the *first* WST application to actual galaxy observations. In particular, we use the WST to analyze galaxy observations from the twelfth data release (DR12) [69] of the Baryon Oscillation Spectroscopic Survey (BOSS), a part of Sloan Digital Sky Survey, SDSS-III [70,71], and in particular the CMASS sample. We include the effects of redshift-space anisotropy, nontrivial survey geometry, the shortcomings of the dataset through a set of systematic weights and the Alcock-Paczynski effect, following the commonly adopted steps for the power spectrum analysis. In order to model the cosmological dependence of the WST coefficients extracted from galaxy observations, we make use of state-of-the-art simulated mocks that have been tuned to match the clustering properties of CMASS, through a set of free parameters absorbing the physics of galaxy formation, that we marginalize over. Using this framework, we perform a likelihood analysis of the BOSS CMASS dataset, which allows us to infer the values of cosmological parameters with the WST estimator. We finally compare our WST results against the ones obtained by the multipoles of the regular galaxy power spectrum, which we use as a reference, and clarify certain approximations in our analysis.

The rest of the paper is structured as follows: in Sec. II we introduce the wavelet scattering transform, and in Sec. III we lay out all the details related to the analysis of the BOSS dataset, and of the associated mocks, using the WST and the power spectrum multipoles. We then present our results in Sec. IV, while concluding in Sec. V. More technical results are included in Appendices A, B, C, and D.

## II. WAVELET SCATTERING TRANSFORM

The wavelet scattering transform estimator [47,53] was originally proposed in the context of signal processing in computer vision as a means of capturing the statistical properties of an input field. In addition to exhibiting a set of powerful and well-understood mathematical properties (superseding the ones of the conventional power spectrum) [53], it was also shown to provide key insights into the nature of convolutional neural networks [47]. As a result, and as we will see more clearly below, it can constitute an ideal middle-ground between these two types of approaches.

In the WST approach, a given input field,  $I(\mathbf{x})$ , is subjected to two basic nonlinear operations: wavelet convolution and modulus. That is, if  $\psi_{j_1, l_1}(\mathbf{x})$  is an oriented wavelet probing a scale  $j_1$  and angle  $l_1$ , under the fundamental WST operation,  $I(\mathbf{x})$  will be transformed as:

$$I'(\mathbf{x}) = |I(\mathbf{x}) * \psi_{j_1, l_1}(\mathbf{x})|, \quad (1)$$

where  $*$  denotes convolution. Taking the expectation value<sup>1</sup> of Eq. (1) produces a WST coefficient,  $S$ , which is nothing else than a real number characterizing the field. Combined with a family of localized wavelets  $\psi_{j_1, l_1}(\mathbf{x})$  probing a range of scales  $j_1$  and angles  $l_1$ , successive applications of the above procedure give rise to a *scattering network*, the WST coefficients,  $S_n$ , of which are given by the following relations up to order  $n = 2$ :

$$\begin{aligned} S_0 &= \langle |I(\mathbf{x})| \rangle, \\ S_1(j_1, l_1) &= \langle |I(\mathbf{x}) * \psi_{j_1, l_1}(\mathbf{x})| \rangle, \\ S_2(j_2, l_2, j_1, l_1) &= \langle |(|I(\mathbf{x}) * \psi_{j_1, l_1}(\mathbf{x})|) * \psi_{j_2, l_2}(\mathbf{x})| \rangle, \end{aligned} \quad (2)$$

where in Eq. (2) and hereafter the angular brackets,  $\langle \cdot \rangle$ , denote averaging over the sample. Given that a convolution with an oriented wavelet reflects spatial and angular information about the input field, it can be intuitively understood how the WST coefficients of order  $n$  encode clustering information analogous to the  $2^n$ -point correlation function [53,55]. As such, the network in Eq. (2) generates a compact basis of coefficients that can be used to partially characterize the higher-order clustering properties of a physical field, a task of particular interest to modern cosmology. Furthermore, the fundamental WST operations, namely convolution and modulus, together with its hierarchical architecture resemble a CNN with fixed kernels [47,53]. The combination of the above properties leads to a powerful estimator that can capture the non-Gaussian information content encoded in a physical field, similar to a CNN, but while retaining the desired interpretability of conventional clustering statistics (e.g., correlation function) through a basis of a few WST coefficients. As opposed to the regular clustering statistics, the WST can extract higher-order correlations without raising the target field to high powers [67], which is computationally more efficient. Lastly, the WST has demonstrated the ability to better access the information content carried in physical fields with heavy-tailed probability distributions [67], a case that is particularly challenging for higher-order moments to describe [14]. A pedagogical overview of various other properties of the WST (such as texture characterization or field generation) can be found in Ref. [67], whereas the

<sup>1</sup>In practice, this corresponds to taking the spatial average of the field.

formal mathematical description is discussed in detail in Refs. [47,53–56].

The relations in Eq. (2) can be generalized to allow for operations on a target field raised to a given power,  $q$ , in the following manner:

$$\begin{aligned} S_0 &= \langle |I(\mathbf{x})|^q \rangle, \\ S_1(j_1, l_1) &= \langle |I(\mathbf{x}) * \psi_{j_1, l_1}(\mathbf{x})|^q \rangle, \\ S_2(j_2, l_2, j_1, l_1) &= \langle |(|I(\mathbf{x}) * \psi_{j_1, l_1}(\mathbf{x})|)^q * \psi_{j_2, l_2}(\mathbf{x})|^q \rangle, \end{aligned} \quad (3)$$

where the choice of values of  $q < 1$  or  $q > 1$  emphasizes underdense or overdense regions, respectively, and  $q = 1$  recovers the basic WST case. In the first 3D WST application to the large-scale structure of the universe [52], it was shown that highlighting cosmic voids with values of  $q < 1$  led to a substantial increase in the information extracted on fundamental parameters, particularly the sum of the neutrino masses, matching and also exceeding the performance of the marked power spectrum [28,31]. In order to leverage this property, we choose to work with the relations in Eq. (3), a choice also adopted by the 3D molecular chemistry WST application of [65].

The input WST field  $I(\mathbf{x})$ , that we will specify in the next section, can have an arbitrary number of dimensions, as far as the WST is concerned. In the context of the 3D LSS observations we will focus on in this work,  $I(\mathbf{x})$  will be a 3D field. In general, a family of wavelets,  $\psi_{j_1, l_1}(\mathbf{x})$ , can be generated by performing dilations and rotations on a mother wavelet, which can similarly take various forms according to the desired application. In the 3D WST implementation of this analysis,<sup>2</sup> the mother wavelet is a solid harmonic, multiplied by a Gaussian envelope, of the form

$$\psi_l^m(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}} e^{-|\mathbf{x}|^2/2\sigma^2} |\mathbf{x}|^l Y_l^m\left(\frac{\mathbf{x}}{|\mathbf{x}|}\right), \quad (4)$$

where  $Y_l^m$  are the familiar Laplacian spherical harmonics and  $\sigma$  is the Gaussian width in units of the field pixels. The dilations are then described by the following rescaling of the wavelet argument:

$$\psi_{j_1, l_1}^m(\mathbf{x}) = 2^{-3j_1} \psi_{l_1}^m(2^{-j_1} \mathbf{x}). \quad (5)$$

If we sum over the index  $m$ , and consider  $l$  to describe the angular information of the wavelet family, then the WST coefficients in this particular case will be given by:

<sup>2</sup>This was first introduced in the context of molecular chemistry applications.

$$\begin{aligned}
 S_0 &= \langle |I(\mathbf{x})|^q \rangle, \\
 S_1(j_1, l_1) &= \left\langle \left( \sum_{m=-l_1}^{m=l_1} |I(\mathbf{x}) * \psi_{j_1, l_1}^m(\mathbf{x})|^2 \right)^{\frac{q}{2}} \right\rangle, \\
 S_2(j_2, j_1, l_1) &= \left\langle \left( \sum_{m=1}^{m=l_1} |U_1(j_1, l_1)(\mathbf{x}) * \psi_{j_2, l_1}^m(\mathbf{x})|^2 \right)^{\frac{q}{2}} \right\rangle, \quad (6)
 \end{aligned}$$

with

$$U_1(j_1, l_1)(\mathbf{x}) = \left( \sum_{m=-l_1}^{m=l_1} |I(\mathbf{x}) * \psi_{j_1, l_1}^m(\mathbf{x})|^2 \right)^{\frac{1}{2}}. \quad (7)$$

After performing the 1st order convolution of the input field over a scale determined by  $j_1$  in Eq. (6), any information in scales smaller than that will be obscured in any subsequent convolutions of said field within the scattering network. Indeed, the 2D weak lensing (WL) application by Ref. [60] found that  $S_2$  coefficients with  $j_2 < j_1$  did not contribute any substantial cosmological information.<sup>3</sup> As a result, we also choose to work with  $S_2$  coefficients with only  $j_2 > j_1$ , as we also did in [52]. We additionally point out that, as opposed to Eq. (3), the particular implementation (6) developed by Ref. [65], that we follow, uses the same angular scale  $l_2 = l_1$  for the second order coefficient  $S_2 \equiv S_2(j_2, j_1, l_1)$ . Even though such a restriction most likely causes some loss of angular information, it significantly reduces the associated computational cost of the 3D WST evaluations, a tradeoff that was found to still perform very well in the 3D LSS application of [52].

To summarize, given an input 3D field with a resolution of NGRID cells on a side, a number of total spatial dyadic scales  $J$  [that can never exceed  $\log_2(\text{NGRID})$ ] and total orientations  $L$ , the indices

$$(j, l) \in ([0, \dots, J-1, J], [0, \dots, L-1, L]), \quad (8)$$

give rise to a total of

$$S_0 + S_1 + S_2 = 1 + (L+1)(J^2 + 3J + 2)/2 \quad (9)$$

WST coefficients up to 2nd order. The final choices that need to be determined for a WST evaluation are the values of the power  $q$  and Gaussian width  $\sigma$ , that we will appropriately choose in the next section.

We should note, at this point, that when working with isotropic input fields, dimensionality reduction techniques can further reduce the number of WST coefficients to work with. These include averaging over all  $l$  orientations for a

given spatial scale  $j$ , in order to construct isotropic coefficients [57,58,60] or less aggressive reduction techniques that aim to retain a larger degree of isotropy [72]. Given that in this work we will apply the WST on an *anisotropic* physical field (as is the one determined by galaxy observations in redshift space), we will not consider this reduction. Lastly, we note that the WST coefficients are commonly normalized as follows:

$$\begin{aligned}
 \bar{S}_0 &= \log(S_0), \\
 \bar{S}_1 &= \log(S_1/S_0), \\
 \bar{S}_2 &= \log(S_2/S_1), \quad (10)
 \end{aligned}$$

a choice adopted by several past applications [55,57,58,60,61]. Despite the fact that the WL WST applications on 2D shear maps [60,61] found this renormalized basis to break degeneracies between  $\Omega_m$  and  $\sigma_8$ , our previous 3D application to cosmological density fields of Ref. [52] did not find any noticeable gains in information associated with this basis. As a consequence, we choose to work with the bare WST coefficients given by Eq. (6).

### III. ANALYSIS

In this section, we lay out the details of the particular galaxy dataset and mocks in our analysis, as well as of the procedure we follow in order to extract the WST and power spectrum estimators out of them in each case.

#### A. Dataset

This work uses galaxy observations from the twelfth data release (DR12) [69] of the BOSS,<sup>4</sup> a part of Sloan Digital Sky Survey, SDSS-III [70,71]. Specifically, we work with CMASS data in the redshift range  $0.46 < z < 0.60$  that were observed from two separate parts of the sky, the North (NGC) and South Galactic Cap (SGC). In light of the fact that these are two distinct subsets of observations, we hereafter evaluate our clustering statistics on each one of them separately and then obtain the average, weighted by the corresponding values of their angular footprint, following the standard procedure in past analyses of BOSS data [73,74]. To be specific, if  $X_{\text{NGC}}$  and  $X_{\text{SGC}}$  is our statistic (be it WST or  $P(k)$  multipoles) evaluated from the NGC and SGC parts, respectively, then the resulting data vector used in our analysis is always given by:

$$X_{\text{N+S}} = \frac{(A_{\text{NGC}}X_{\text{NGC}} + A_{\text{SGC}}X_{\text{SGC}})}{(A_{\text{NGC}} + A_{\text{SGC}})}, \quad (11)$$

<sup>3</sup>We do note, nevertheless, that using equivariant wavelets on 2D fields the work of Ref. [72] did find some residual power to be carried in those usually discarded coefficients.

<sup>4</sup>All BOSS data, as well as the accompanied covariance PATCHY mocks, are publicly available at <https://data.sdss.org/sas/dr12/boss/lss/>.



where  $A_{\text{NGC}} = 6851 \text{ deg}^2$  and  $A_{\text{SGC}} = 2525 \text{ deg}^2$ . We should note, at this point, that even though the redshift range of the entire CMASS sample commonly adopted in BOSS analyses is actually  $0.43 < z < 0.70$ , we work with a narrower  $z$  cut because this was the one used for the production of our galaxy mocks, as we will see in the next section. Due to the same reason, we only work with the CMASS, rather than also with the LOWZ BOSS sample, even though our WST estimator can be straightforwardly applied to any set of galaxy observations.

Each dataset is accompanied by a corresponding random unclustered catalog, with the exact same angular footprint and selection function, in order to enable the evaluation of clustering statistics (as we will see below). We choose to work with the random catalog that has a number density  $50\times$  greater than the one of the corresponding observed samples, a choice commonly adopted in previous analyses [73,74].

We now proceed to explain the procedure followed to evaluate the fractional overdensity field from the BOSS dataset, which serves as the fundamental quantity of interest needed to extract both the WST and the P(k) statistic. To do so, we start by converting the observed galaxy sky coordinates, right ascension (RA), declination (DEC) and redshift  $z$  into comoving Cartesian coordinates,  $x, y, z$ , always assuming a cosmology of  $\Omega_m = 0.3152$ ,  $h = 0.6736$ , which, as we will see below, will correspond to our chosen fiducial cosmology. We account for the potential errors introduced when assuming an incorrect cosmology to perform this conversion, which are known as the Alcock-Paczynski (AP) distortion [75]. We will explain our strategy in Sec. III C for the power spectrum and in Sec. III D for the WST. Using the publicly available package `nboddykit`,<sup>5</sup> we then proceed to embed the sample into a cubic box with a comoving side equal to  $L = 2820 \text{ Mpc}/h$ , which corresponds to the smallest possible cube that can embrace our (irregularly shaped) sample. Finally, the desired quantity is the (weighted) fractional overdensity field of data in a realistic survey format, commonly referred to as the Feldman-Kaiser-Peacock (FKP) field,  $F(r)$  [24], which we evaluate on a mesh through the following relationship:

$$F(\mathbf{r}) = \frac{w_{\text{FKP}}(\mathbf{r})}{I_2^{1/2}} [w_c(\mathbf{r})n_g(\mathbf{r}) - \alpha_r n_s(\mathbf{r})]. \quad (12)$$

Here  $n_g(\mathbf{r})$  and  $n_s(\mathbf{r})$  are the observed number density of the galaxies and the objects of the random catalog, respectively, and  $\alpha_r$  denotes the ratio between the (weighted) total number of objects in the galaxy catalog over the corresponding value of the synthetic random one. The BOSS dataset further includes 3 weights that reflect the incompleteness of the observed sample: a redshift failure

weight,  $w_{\text{rf}}$ , a fiber collision weight,  $w_{\text{fc}}$  and a systematics weight,  $w_{\text{sys}}$ . They enter Eq. (12) as a combined contribution [73,74]

$$w_c(\mathbf{r}) = (w_{\text{rf}}(\mathbf{r}) + w_{\text{fc}}(\mathbf{r}) - 1.0)w_{\text{sys}}(\mathbf{r}). \quad (13)$$

The remaining weight in Eq. (12) is the FKP weight [24],

$$w_{\text{FKP}}(\mathbf{r}) = [1 + \bar{n}_g(\mathbf{r})P_0]^{-1}, \quad (14)$$

for  $P_0 = 10^{-4} \text{ Mpc}^3/h^3$ , which is meant to ensure optimal extraction of information at small scales and is also provided for each galaxy of the sample. Finally, the normalization factor

$$I_2 = \int d^3\mathbf{r} w_{\text{FKP}}^2(\mathbf{r}) \langle w_c(\mathbf{r})n_g(\mathbf{r}) \rangle^2 \quad (15)$$

is meant to normalize the amplitude of the power spectrum with respect to the observed power in an instance of no survey selection. It is straightforward to see that in the absence of a weighting scheme, Eq. (12) gives the regular galaxy overdensity field evaluated from a sample. The FKP field from Eq. (12) is the fundamental quantity of interest to extract from the data, which, as we will see below, can either be fed into Eq. (6) to obtain the observed WST coefficients or get Fourier transformed in order to obtain the multipoles of the anisotropic galaxy power spectrum.

It should be noted, at this point, that the weighting scheme (13) was designed to account for the impact of the dataset incompleteness on the power spectrum, rather than the WST. As a result, it is possible that a different set of weights is needed in order to fully account for these effects on the WST coefficients. Given however, that the WST (partly) contains clustering information comparable to that in the 2-point correlation function, we expect Eq. (13) to at least partially capture the shortcomings of the dataset, from a WST point of view, and consider it a reasonable starting point for this first WST application. We defer a more detailed investigation of how to optimally model this effect for the WST to future work. Likewise, even though the FKP weights from Eq. (14) were designed to ensure the optimal recovery of information by the power spectrum, we include them in the WST analysis as well, in order to maintain consistent inputs across our pipeline.

## B. Mocks

Previous analyses of BOSS data use more traditional estimators, which mostly relied on perturbation theory models of various kinds to capture the cosmological dependence of the target clustering statistics (with a few representative examples being [20,73,74,76–78]). On the other hand, given the lack of a first principles theory model in place to predict its cosmological dependence, the WST approach demands the use of a full set of simulated mocks.

<sup>5</sup><https://nboddykit.readthedocs.io/en/latest/index.html>.

In this section we introduce the set of mocks we will use for our model and covariance matrix for the WST (and the power spectrum, for comparison) in our final likelihood analysis.

### I. ABACUSSUMMIT mocks

In order to model the cosmological dependence of our estimators, we use the publicly available suite of ABACUSSUMMIT simulations [79],<sup>6</sup> which were performed using the state-of-the-art ABACUS N-body code [80,81]. Having evolved  $6912^3$  dark matter particles in a cubic box of a side equal to  $L_{\text{Box}} = 2000 \text{ Mpc}/h$  (for the “base” configuration), corresponding to a particle mass resolution  $M_p = 2.1 \times 10^9 M_\odot/h$ ,<sup>7</sup> these simulations are able to not only match but also exceed the requirements of DESI [9], making them the best available set of simulations to work with. The gravitationally bound dark matter halos in the ABACUSSUMMIT collection were identified using the state-of-the-art spherical overdensity halo finder, COMPASO [82].

We will now briefly summarize the cosmological landscape of the ABACUSSUMMIT, starting with the fiducial base configuration, which corresponds to the mean values of the “base\_plikHM\_TTT\_EEE\_lowl\_lowE\_lensing” version of the *Planck* 2018 [83]  $\Lambda$ CDM cosmology. In order to enable averaging over the effects of cosmic variance, the base cosmology was also run using 24 additional different random initial phases, in addition to the base box. We begin by considering constraints around 4 cosmological parameters,  $\{\omega_b, \omega_c, n_s, \sigma_8\} = \{0.02237, 0.120, 0.9649, 0.8114\}$ , where  $\omega_X = \Omega_X h^2$  and  $\sigma_8$  refers to the amplitude of density fluctuations traced by the combination of cold dark matter and baryons, “cb”, in the presence of massive neutrinos with  $\omega_\nu = 0.0006442$ . To capture the cosmological dependence on the above 4 parameters, we further consider additional simulations which vary each one of the parameters, in turn, and in a step-wise fashion, away from the fiducial background while keeping the rest fixed. This enables the evaluation of first-order derivatives. The exact list (as well as the associated parameter variations of these ‘first-order derivative grid’ cosmologies) are listed in Table I, while we also add that these were phase-matched to the base box, in order to cancel out the effects of cosmic variance upon taking central differences. All simulations have kept the value of the angular size of the sound horizon at last scattering,  $\theta_*$ , fixed to the corresponding value derived from measurements by the *Planck* satellite [83]:  $100\theta_* = 1.041533$ , which implies a corresponding value of  $h = 0.6736$ , for the base cosmology. Even though the Hubble constant was

TABLE I. A list of all 8 ABACUSSUMMIT first-order derivative cosmologies we use in this work, together with the corresponding values of the 4 cosmological parameters varied. The first row corresponds to the base cosmology, shown for reference.

$\omega_b$	$\omega_c$	$n_s$	$\sigma_8$
0.02237	0.1200	0.9649	0.8114
0.02282	0.1200	0.9649	0.8114
0.02193	0.1200	0.9649	0.8114
0.02237	0.1240	0.9649	0.8114
0.02237	0.1161	0.9649	0.8114
0.02237	0.1200	1.0249	0.8114
0.02237	0.1200	0.9049	0.8114
0.02237	0.1200	0.9649	0.8698
0.02237	0.1200	0.9649	0.7532

not explicitly varied in ABACUSSUMMIT, the choice of a fixed  $\theta_*$  enables the evaluation of  $h$  as an additional derived parameter, an option that we will consider below. The values of all the other cosmological parameters, which we keep fixed, together with a detailed description of the ABACUSSUMMIT simulations can be found in Ref. [79].

In order to generate realistic galaxy mock samples from the underlying dark matter and halo catalogs, ABACUSSUMMIT uses the Halo Occupation Distribution (HOD) framework through a flexible package called ABACUSHOD [84].<sup>8</sup> In the baseline HOD implementation [85], simulated halos host galaxies based on a semi-analytical probabilistic model that depends on 5 parameters. In particular, if  $M_{\text{cut}}$  denotes the minimum mass of a halo that can host a central galaxy,  $\kappa M_{\text{cut}}$  the minimum halo mass to host a satellite galaxy and  $M_1$  characterizes the typical halo mass that hosts one satellite galaxy, the mean expected number of central,  $\bar{n}_{\text{cent}}(M)$ , and satellite,  $\bar{n}_{\text{sat}}(M)$ , galaxies assigned to a halo of mass  $M$  are given by:

$$\bar{n}_{\text{cent}}(M) = 0.5 \operatorname{erfc} \left[ \frac{\log_{10} \left( \frac{M_{\text{cut}}}{M} \right)}{\sqrt{2}\sigma} \right], \quad (16)$$

and

$$\bar{n}_{\text{sat}}(M) = \left[ \frac{M - \kappa M_{\text{cut}}}{M_1} \right]^\alpha \bar{n}_{\text{cent}}(M). \quad (17)$$

The parameters  $\alpha$  and  $\sigma$  calibrate the relations (16) and (17), and fully characterize the standard HOD model. We note that Eqs. (16) and (17) are applicable to luminous red galaxies (LRGs) [86,87], which mostly dominate the CMASS sample that we will work with. Out of a rich

<sup>6</sup>Detailed information on all the simulations, as well as on how to access them, can be found at <https://abacussummit.readthedocs.io/en/latest/index.html>.

<sup>7</sup> $h$  is the dimensionless Hubble constant,  $h = H_0 / (100 \text{ km s}^{-1} \text{ Mpc}^{-1})$ .

<sup>8</sup>The package is publicly available as part of ABACUSUTILS <http://https://github.com/abacusorg/abacusutils>, together with an associated instruction manual at <https://abacusutils.readthedocs.io/en/latest/hod.html>. In particular, we use the ABACUSUTILS version 1.0.4.

variety of extensions going beyond the vanilla HOD framework described above, we will adopt a version that accounts for the effect of the velocity bias of LRGs, a step shown to be necessary for an accurate fit against both BOSS data [84] and hydrodynamical simulations [88]. The two additional HOD parameters it introduces are the central velocity bias,  $\alpha_c$ , which is meant to account for the mismatch between the velocity of central galaxies and the one of halo centers, and the satellite velocity bias,  $\alpha_s$ , which captures the equivalent effect for satellite galaxies. For a more in-depth discussion on more sophisticated HOD parametrizations (such as, for example, including assembly bias [89]) interested readers are referred to the relevant works referenced above.

We can finally proceed to use our adopted 7-parameter HOD framework in order to generate galaxy mocks for our BOSS analysis. In particular, we closely follow the procedure laid out in Refs. [84,90] and search for an ABACUS-derived galaxy mock that best fits the redshift-space 2-point correlation function,  $\xi(r_\perp, r_\parallel)$ , of the BOSS DR12 CMASS sample introduced in Sec. III A, [averaged over North and South according to Eq. (11)], for the base cosmology. Here  $r_\perp, r_\parallel$  are the separations perpendicular and parallel to the line of sight, respectively, and we fit the CMASS correlation function using 8 logarithmic bins in the range 0.169–0.30 Mpc/h for  $r_\perp$  and using 6 linearly spaced bins between 0 and 30 Mpc/h for  $r_\parallel$ , for the redshift cut  $0.46 < z < 0.60$ . The HOD parameters giving the best fit through this procedure, that we take as the fiducial HOD parameters, correspond to the following values:

$$\begin{aligned} & \{\alpha, \alpha_c, \alpha_s, \kappa, \log M_1, \log M_{\text{cut}}, \sigma\} \\ & = \{0.9022, 0.2499, 1.1807, 0.3288, 14.313, \\ & \quad 12.8881, 0.02084\}, \end{aligned} \quad (18)$$

where  $M_{\text{cut}}$  and  $M_1$  are expressed in units of  $M_\odot/h$ . We then proceed to use ABACUSHOD, always with the parameters in Eq. (18) as input, so as to produce CMASS mocks for each one of the 8 first-order derivative cosmologies in Table I, as well as for the 25 boxes of the fiducial cosmology. In order to null out any residual effects of cosmic variance during the evaluation of the derivatives, in addition, we generate 20 different HOD realizations for each one of the 8 derivative cosmologies. Finally, in order to capture the effects of varying the values of the HOD parameters, and in a direct analogy to the procedure followed for the four cosmological parameters, we proceed to generate first-order derivative mocks for HOD variations. We kept all parameters fixed to the fiducial values and successively varied one HOD parameter at a time, in a step-wise fashion. With  $\theta_{\text{HOD}}$  being the vector of HOD parameter values from Eq. (18), we take steps  $\theta_{\text{HOD}} \pm \Delta\theta_{\text{HOD}}$  as follows:

$$\begin{aligned} & \{\Delta\alpha, \Delta\alpha_c, \Delta\alpha_s, \Delta\kappa, \Delta \log M_1, \Delta \log M_{\text{cut}}, \Delta\sigma\} \\ & = \{0.36088, 0.02499, 0.11807, 0.2959, 0.7157, \\ & \quad 0.3866, 0.020\}. \end{aligned} \quad (19)$$

The resulting output of the procedure described in this section is a collection of 25 mocks for the base cosmology and a total of  $20 \times 2 \times 11 = 440$  mocks for the cosmological derivative variations, which constitute the full set of simulations that we will use in our analysis.

## 2. PATCHY mocks

In order to construct a Gaussian likelihood for our inference framework, that we will explain in detail in Sec. III E, we also need to construct an accurate covariance matrix in addition to the model for the observable. To do so, we use the 2048 realizations of the publicly available MULTIDARK-PATCHY mocks<sup>9</sup> [87,91], hereafter referred to as PATCHY mocks, a collection large enough to make them ideal candidates for the evaluation of a properly converged covariance matrix. These mocks were produced through a hybrid combination of an approximate gravity solver and a reference simulation [92] that evolved 3840<sup>3</sup> dark matter particles on a cubic volume of side 2.5 Gpc/h, using the code GADGET-2 [93], with a baseline cosmology given by  $\{\Omega_b, \Omega_m, n_s, \sigma_8, h\} = \{0.0482, 0.307, 0.961, 0.829, 0.6778\}$ . Gravitationally bound halos were identified using the bound density maximum halo finder [94], which were subsequently populated with galaxies using the halo abundance matching technique (HAM) [95], an alternative to the HOD method described above. The PATCHY mocks were finally shaped into the realistic survey geometry of the BOSS CMASS dataset, also split into the separate NGC and SGC observed parts of the sky, while each galaxy was assigned a set of systematic weights [similarly to Eq. (13)]

$$w_c(\mathbf{r}) = w_{\text{fc}}(\mathbf{r})w_{\text{veto}}(\mathbf{r}). \quad (20)$$

The weights include fiber collisions,  $w_{\text{fc}}$ , and a veto mask,  $w_{\text{veto}}$ , capturing the rest of the associated shortcomings of the dataset. The galaxies are also assigned FKP weights, according to Eq. (14). Since the PATCHY mocks were cast into a survey format, we treat them as the data and repeat the exact same procedure detailed in Sec. III A in order to generate the resulting FKP field from Eq. (12), but using the weighting scheme in Eq. (20) rather than in Eq. (13).

We will assume a cosmology-independent covariance matrix [96,97] and combine the PATCHY mocks with the ABACUSSUMMIT suite, even though strictly speaking the two sets correspond to different fiducial cosmologies and used different mock-generating procedures. We also note

<sup>9</sup>Available at [https://data.sdss.org/sas/dr12/boos/lss/dr12\\_multidark\\_patchy\\_mocks/](https://data.sdss.org/sas/dr12/boos/lss/dr12_multidark_patchy_mocks/).



that mixing different ways of modeling the theory vector and the covariance matrix is common practice in BOSS analyses (e.g., [20,73,74,76–78]). Under this approximation, we use the ABACUSSUMMIT fiducial cosmology (rather than the one of PATCHY) to convert the galaxy positions from sky coordinates RA, DEC, and  $z$  into comoving Cartesian ones. We finally add that the PATCHY mocks are accompanied by their own set of randomly generated catalogs, both for the NGC and the SGC, each containing  $\sim 50\times$  the number of objects in the corresponding actual galaxy mock.

## C. Power spectrum

In this section we explain how we evaluate the multipoles of the anisotropic redshift-space power spectrum, the predictions of which we will use as a reference to assess the performance of the WST. Given that the BOSS data (as well as the PATCHY mocks) and the ABACUSSUMMIT mocks come in a different format, we will follow a different strategy to extract the power spectrum multipoles in each case.

### 1. BOSS data and PATCHY mocks

When working with galaxy data or mocks in a nontrivial survey geometry, with a corresponding FKP field from Eq. (12), the multipoles of the anisotropic redshift-space power spectrum,  $\hat{P}_\ell(k)$ , can be evaluated using the “Yamamoto” estimator [24,98–100]:

$$\hat{P}_\ell(k) = \frac{(2\ell + 1)}{I_2} \int \frac{d\Omega_k}{4\pi} \left[ \int d^3\mathbf{r}_1 F(\mathbf{r}_1) e^{i\mathbf{k}\cdot\mathbf{r}_1} \times \int d^3\mathbf{r}_2 F(\mathbf{r}_2) e^{i\mathbf{k}\cdot\mathbf{r}_2} L_\ell(\hat{\mathbf{k}} \cdot \hat{\mathbf{r}}_2) - P_\ell^{\text{shot}}(\mathbf{k}) \right]. \quad (21)$$

Here  $L_\ell$  is the Legendre polynomial of order  $\ell$ ,  $d\Omega_k$  is the differential solid angle element in Fourier space and the term  $P_\ell^{\text{shot}}$  is the shot noise contribution:

$$P_\ell^{\text{shot}}(\mathbf{k}) = (1 + \alpha_r) \int d^3\mathbf{r} w^2(\mathbf{r}) \bar{n}_g(\mathbf{r}) L_\ell(\hat{\mathbf{k}} \cdot \hat{\mathbf{r}}), \quad (22)$$

which vanishes for higher order multipoles  $\ell > 0$ , and where  $w(\mathbf{r}) = w_c(\mathbf{r})w_{\text{FKP}}(\mathbf{r})$ . We use Eqs. (21) and (22) to extract the galaxy power spectrum multipoles from the BOSS catalogs and the PATCHY mocks, both of which are shaped into a realistic survey geometry. We do so using `nbodykit`, which follows the optimized fast Fourier transform (FFT)-based implementation of Eq. (21), developed by Ref. [101]. In particular, we use Eq. (12) to evaluate the FKP field using the triangular shaped cloud (TSC) mass assignment scheme [102] on `NGRID` = 500 cubic cells on the side, which corresponds to a high enough resolution for an accurate description of the scales we will consider. We then evaluate the first 2 nonvanishing

multipoles,  $\ell = \{0, 2\}$ , of the power spectrum through Eq. (21) using 46 linearly spaced bins of width  $\Delta k = 0.01$  h/Mpc within the  $k$  range 0.001–0.50 h/Mpc. Our chosen  $\Delta k$  value has been found to be adequate for recovering the information encoded in the power spectrum [103]. We finally discard wavenumbers larger than  $k_{\text{max}} = 0.25$  h/Mpc, a choice that both matches the ones of previous BOSS analyses [20,73,74,76–78] and also guarantees that the power spectrum and WST both reach a similar minimum scale, such that we perform a fair comparison between the two estimators.

### 2. ABACUSSUMMIT mocks

The ABACUSSUMMIT mocks we work with were produced, by design, in the regular ABACUS periodic cubic geometry with  $L_{\text{Box}} = 2000$  Mpc/h, with the redshift-space anisotropy applied along their Cartesian  $\hat{\mathbf{z}}$ -axis and without any systematic weights applied to them. As a result, their corresponding power spectrum multipoles can be straightforwardly evaluated using standard FFT-based algorithms on a periodic box, which are also supported by `nbodykit`. We do so using a TSC mass assignment scheme with `NGRID` = 1700 cells on the side and the same  $k$ -binning strategy adopted for the data in Sec. III C 1, a set of choices that was once again confirmed to guarantee sufficient accuracy for the range of scales included in our analysis.

However, we should be careful when comparing power spectra evaluated from data on a survey geometry against predictions obtained from periodic cubic mocks. This is because of the fact that, if  $\delta(\mathbf{x})$  is the regular galaxy density field obtained from a periodic volume, and  $W(\mathbf{x})$  the survey window function (that includes both the survey geometry and systematic weights), then a survey will observe a masked density field,  $\delta'(\mathbf{x})$ , given by [24,104]:

$$\delta'(\mathbf{x}) = \delta(\mathbf{x})W(\mathbf{x}), \quad (23)$$

which corresponds to a convolution in Fourier space. The FKP field of Eq. (12) is such an example. As a consequence, if  $P(\mathbf{k})$  is the theory power spectrum corresponding to  $\delta(\mathbf{x})$  (estimated, for example, from the ABACUSSUMMIT mocks or a perturbation theory model) then the power spectrum,  $\hat{P}(\mathbf{k})$ , observed from a survey with nontrivial geometry will be modified as follows [24,73,74,104,105]:

$$\hat{P}(\mathbf{k}) = \int \frac{d^3\mathbf{k}'}{(2\pi)^3} P(\mathbf{k}') |\tilde{W}(\mathbf{k} - \mathbf{k}')|^2, \quad (24)$$

where

$$\tilde{W}(\mathbf{k}) = \frac{\alpha_r}{I_2^{1/2}} \int d^3\mathbf{r} \bar{n}_s(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}}, \quad (25)$$



is the Fourier transform (FT) of the window function, that obeys the normalization

$$\int \frac{d^3\mathbf{k}}{(2\pi)^3} |\tilde{W}(\mathbf{k})|^2 = 1. \quad (26)$$

To evaluate the multipoles,  $\hat{P}_\ell(k)$  in Eq. (24), we transform the ABACUS-derived power spectrum multipoles,  $P_\ell(k)$ , to get the corresponding correlation function multipoles,  $\xi_\ell(s)$ , as follows [74,104]<sup>10</sup>:

$$\xi_\ell(s) = i^\ell \int \frac{dk k^2}{2\pi^2} j_\ell(ks) P_\ell(k), \quad (27)$$

with  $j_\ell(ks)$  the spherical Bessel functions of order  $\ell$ . Given the configuration space multipoles of Eq. (25),  $W_\ell^2(s)$ , we can modify the correlation function multipoles to account for this effect. Explicitly, the window-corrected multipoles,  $\hat{\xi}_\ell(s)$ , will be given (up to  $\ell = 2$ ) by [74]:

$$\begin{aligned} \hat{\xi}_0(s) &= \xi_0(s) W_0^2(s) + \frac{1}{5} \xi_2(s) W_2^2(s), \\ \hat{\xi}_2(s) &= \xi_0(s) W_2^2(s) + \xi_2(s) \left[ W_0^2(s) + \frac{2}{7} W_2^2(s) \right]. \end{aligned} \quad (28)$$

Finally, these can be transformed back to the Fourier space, in order to give the window-corrected multipoles,  $\hat{P}_\ell(k)$ , through:

$$\hat{P}_\ell(k) = (-i)^\ell 4\pi \int ds s^2 j_\ell(ks) \hat{\xi}_\ell(s). \quad (29)$$

We use Eq. (29) to evaluate the window-corrected power spectrum multipoles from the ABACUSSUMMIT mocks, to be compared against the corresponding results from the BOSS data and the PATCHY mocks.<sup>11</sup> To apply Eq. (28), we use the publicly available results for  $W_\ell^2(s)$  provided by Ref. [74] (which evaluates these functions with the pair-counting method proposed by Ref. [104]), for both the NGC and the SGC patches of the BOSS CMASS sample, separately. Following the standard practice in previous BOSS analyses [73,74,105], we average over the window contributions for the North and the South, according to Eq. (11). An alternative way to handle the effects of survey geometry would be to deconvolve the data, instead, as more recently proposed by Ref. [107].

Furthermore, as we discussed in Sec. III A, the assumption of a given (and potentially incorrect) cosmology when converting the data (and the PATCHY mocks)

from sky coordinates, RA, DEC and  $z$ , into comoving coordinates might introduce an error in our analysis [75]. To account for this AP effect on the estimation of the power spectra from the mocks, we work as follows [108]: if  $r_\parallel$  and  $r_\perp$  are the mock galaxy coordinates parallel and perpendicular to the line of sight (which coincides with the Cartesian  $\hat{z}$  direction for the cubic mocks), they should then be rescaled according to the following relations:

$$r_{\parallel,\text{ref}} = r_{\parallel,\text{sim}} \left( \frac{H_{\text{sim}}(z) H_{0,\text{ref}}}{H_{0,\text{sim}} H_{\text{ref}}(z)} \right), \quad (30)$$

and

$$r_{\perp,\text{ref}} = r_{\perp,\text{sim}} \frac{d_{A,\text{ref}}(z)}{d_{A,\text{sim}}(z)}. \quad (31)$$

The subscripts ‘‘sim’’ and ‘‘ref’’ in Eqs. (30) and (31) denote the true cosmology of each mock and the reference cosmology assumed for the conversion (corresponding to the fiducial, in our case), respectively, and  $d_A$  is the comoving angular diameter distance. That is, before we apply Eqs. (23)–(29) to evaluate the power spectrum multipoles, the galaxy coordinates of each mock are first re-scaled according to Eqs. (30) and (31). Alternatively, one could use an analytical prediction of the AP effect on the anisotropic power spectra (as, for example, done in [20,76–78,109,110]), a procedure that is equivalent to the one described above.

Finally, in Fig. 1 we show the power spectrum multipoles for the base cosmology, as obtained from the ABACUSSUMMIT mocks using Eq. (29) and for the BOSS CMASS dataset, as well as the PATCHY mocks, using Eq. (21). We highlight the excellent monopole agreement between the ABACUSSUMMIT prediction and the corresponding one from the data, down to scales smaller than the ones we work with in this analysis.

## D. WST

To evaluate the WST coefficients for all datasets and mocks used in our analysis, we make use of the publicly available package KYMATIO [111],<sup>12</sup> which implements the WST Eq. (6) for an input 3D density field  $I(\mathbf{x})$ .<sup>13</sup> We evaluate all input density fields on a cubic box of side  $L_{\text{box}} = 2820$  Mpc/h, with  $\text{NGRID} = 282$  grids on the side and the TSC mass assignment scheme, while always adopting the choices  $J = 4, L = 4, \sigma = 0.8$  and  $q = 0.8$ . The choice of a grid cell with a side equal to 10 Mpc/h guarantees that the WST uses information from a minimum

<sup>10</sup>We use the package `mcfit`, which is publicly available at <https://github.com/eelregit/mcfit> and implements the FFTLOG algorithm [106].

<sup>11</sup>We use the corrected normalization coefficients for the power spectrum, as explained in detail in Ref. [107].

<sup>12</sup><https://www.kymatio.io/>.

<sup>13</sup>Strictly speaking, KYMATIO evaluates the sum over all pixels of the input field, rather than the mean, which is the same up to a normalization, and thus equivalent for parameter inference applications.

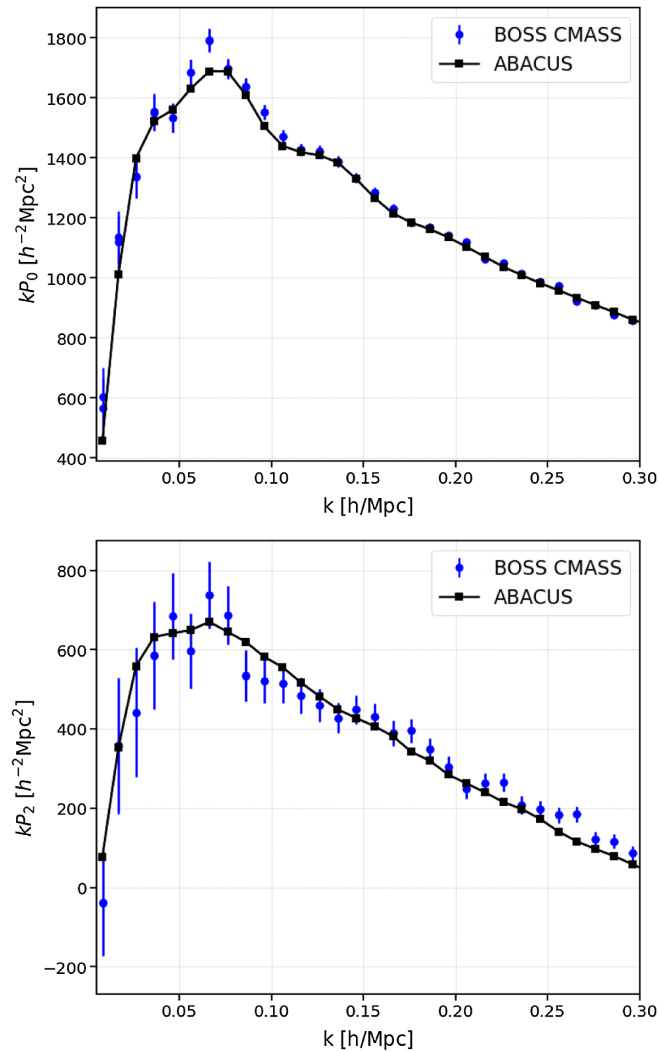


FIG. 1. Redshift-space monopole (top panel) and quadrupole (bottom panel) of the galaxy power spectrum evaluated from the BOSS CMASS dataset (blue circles) and the ABACUSSUMMIT mocks (black squares) for the fiducial cosmology. The  $1\sigma$  error bars on the data have been evaluated from the 2048 PATCHY mock realizations.

scale that is both similar to the one used for the power spectrum multipoles (for which  $k_{\text{max}} = 0.25$  h/Mpc) and also ensures that we do not extract information from a regime that would make the evaluation susceptible to small-scale systematics. This combination corresponds to a basis of  $S_0 + S_1 + S_2 = 76$  total WST coefficients (from Eq. (9)).

### 1. BOSS data and PATCHY mocks

In order to extract the WST coefficients from the data and the PATCHY mocks, we need to apply Eq. (6) with the corresponding FKP field evaluated from Eq. (12), as input. However, we need to be careful at this point, because Eq. (6) assumes a periodic 3D cube as input, rather than a masked density field such as the FKP one. To overcome this obstacle, we modified the public version of KYMATIO

such that Eq. (6), and in particular its fundamental operations of wavelet convolution and modulus, can be used with a masked density field of the form (12) [or more generally (23)] as input (with the technical details discussed in Appendix A). We finally proceed to extract the WST coefficients from the CMASS dataset, as well as from the 2048 PATCHY mocks, using this modified version.

### 2. ABACUSSUMMIT mocks

As we already discussed in Sec. III C 2, the ABACUS-derived mocks from ABACUSHOD have a 3D cubic geometry. However, as we discussed in Sec. III D 1, the WST coefficients extracted from the data were computed from the masked density field of Eq. (12). This implies that, just like in the case of the power spectrum, the effect of survey geometry needs to be taken into consideration in our WST predictions from ABACUS (in fact, the WST analysis is even more sensitive to the survey geometry). In the absence of a model to apply this window correction on the evaluated statistic directly, such as Eq. (29) for the power spectrum, we proceed to cut the ABACUS cubes into the shape of the BOSS CMASS data.<sup>14</sup> For this, we use the public code MAKE\_SURVEY [113],<sup>15</sup> with the exact CMASS angular footprint (for each one of the NGC and the SGC patches) and redshift range, as input. Before feeding the cubic mocks into make\_survey, we first undo the redshift-space distortions effect originally applied along the Cartesian  $\hat{z}$ -axis, such that the code can then implement it along the sky radial direction, resembling the realistic configuration of the actual survey. The necessary galaxy velocities for these RSD manipulations are also provided upon the mock generation by ABACUSHOD. The final resulting output is an equivalent set of galaxy mocks in sky coordinates RA, DEC, and  $z$  that exactly match the 3D geometry of the observed CMASS dataset. To make sure that the clustering properties were not affected during the cut sky implementation, we evaluate the power spectrum multipoles from the new reshaped mocks from Eq. (21) (using, also, an additional set of randoms produced with the same procedure) and confirm that the result matches the one from the corresponding cubic box using Eq. (29), for the fiducial cosmology. One could actually use either of those ways of evaluating the power spectrum, since they are equivalent when handled properly, and which choice to go with is ultimately a matter of preference. Having confirmed the robustness of this procedure, we then proceed to evaluate the FKP fields from all the cut ABACUS mocks using Eq. (12) (but with all the weights set to 1), which we then feed into the set in Eq. (6) so as to finally get the WST coefficients for all cosmologies of Table I.

<sup>14</sup>Alternatively, inpainting techniques could be considered [112].

<sup>15</sup>Available at [https://github.com/mockFactory/make\\_survey](https://github.com/mockFactory/make_survey).

Lastly, as with the power spectrum analysis laid out in Sec. III C, we need to also account for the AP effect introduced by the assumption of a given cosmology when converting the data into comoving coordinates. In the WST case, the mock data were converted (using the true cosmology of each simulation) into sky coordinates upon the cut-sky procedure we described above. To account for the AP effect, we then use the reference fiducial cosmology in order to convert these data back into comoving coordinates, a step that is necessary to evaluate the FKP field from Eq. (12), out of which the WST coefficients will be extracted. This step serves as the WST equivalent to the AP re-scalings applied in Eqs. (30) and (31) for the power spectrum multipoles.

In Fig. 2, we plot the WST coefficients evaluated from the ABACUSSUMMIT mocks, together with the corresponding results from the CMASS data, for the fiducial cosmology. The level of agreement between the ABACUS model and the data, which is even better than the one in the power spectrum case, confirms the validity of our model constructed for the WST.

### E. Likelihood analysis

Having laid out the methodology on how to extract the clustering statistics from both the data and also from the two sets of mocks used in this work, we now proceed to explain our strategy for combining these necessary ingredients into a likelihood analysis of the BOSS dataset. In particular, if  $\mathbf{X}$  is our target estimator (the WST coefficients

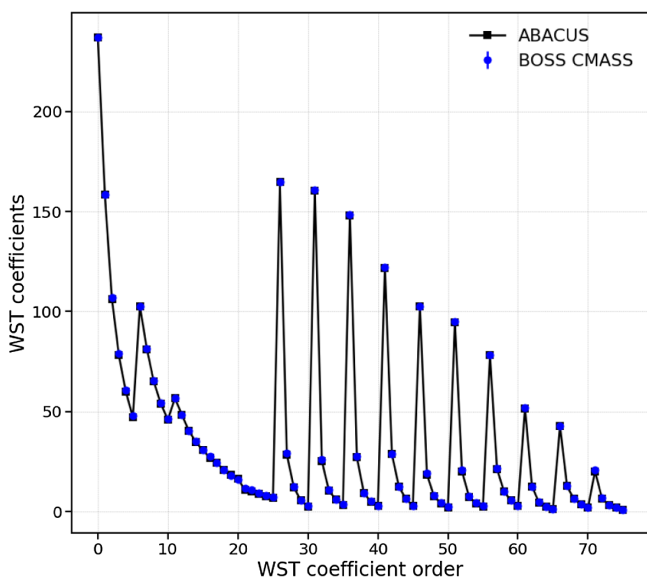


FIG. 2. All 76 WST coefficients evaluated from the BOSS CMASS dataset (blue circles) and the ABACUSSUMMIT mocks (black squares) for the fiducial cosmology. The WST coefficients populate the data vector in order of increasing values of the  $j_1$  and  $l_1$  indices, with the  $l_1$  index varied faster. The  $1\sigma$  error bars on the data (which are too small to be clearly seen on the plot) have been evaluated from the 2048 PATCHY mock realizations.

or the power spectrum multipoles), we assume a Gaussian likelihood,<sup>16</sup>  $\mathcal{L}(\theta|\mathbf{d})$ , given by the following relation:

$$\log \mathcal{L}(\theta|\mathbf{d}) = -\frac{1}{2}[\mathbf{X}_{\mathbf{d}} - \mathbf{X}_l(\theta)]^T C^{-1}[\mathbf{X}_{\mathbf{d}} - \mathbf{X}_l(\theta)] + \text{const.}, \quad (32)$$

where  $\theta$  is the parameter we want to extract from the data  $\mathbf{d}$  and  $\mathbf{X}_{\mathbf{d}}$  is the value extracted from the BOSS data  $\mathbf{d}$ . The covariance matrix  $C$  is estimated from the  $N_{\text{mocks}} = 2048$  PATCHY mocks:

$$C = \frac{1}{N_{\text{mocks}} - 1} \sum_{k=1}^{N_{\text{mocks}}} (\mathbf{X}_p^k - \bar{\mathbf{X}}_p)(\mathbf{X}_p^k - \bar{\mathbf{X}}_p)^T, \quad (33)$$

with  $\bar{\mathbf{X}}_p$  the mean prediction from the  $N_{\text{mocks}}$ . In order to de-bias our prediction for the inverse covariance matrix,  $C^{-1}$ , we apply the Hartlap correction factor [115], as follows:

$$\hat{C}^{-1} = \frac{N_{\text{mocks}} - N_{\mathbf{d}} - 2}{N_{\text{mocks}} - 1} C^{-1}, \quad (34)$$

where  $N_{\mathbf{d}} = 58$  when working with the  $l = 0, 2$  multipoles of the galaxy power spectrum (up to  $k_{\text{max}} = 0.25$  h/Mpc) and  $N_{\mathbf{d}} = 76$  for the WST coefficients as the data vector. Prior to inversion, we make sure that the covariance matrices for both estimators are well-conditioned and can thus be safely inverted in order to be used in the likelihood in Eq. (32). The convergence of the WST covariance is confirmed in Appendix B. The correlation matrix,  $C_{ij}/(C_{ii}C_{jj})$ , of the WST coefficients is shown in Fig. 3, while the corresponding matrix for the power spectrum multipoles is presented in Fig. 4, both evaluated at the fiducial cosmology.

Crucially, the remaining quantity to determine in Eq. (32) is the model,  $\mathbf{X}_l(\theta)$ , which captures the dependence of each estimator on the target set of cosmological parameters,  $\theta$ . In our case this is an 11-dimensional vector consisting of the 4 cosmological parameters varied in Table I and the 7 nuisance parameters of the HOD model, from Eq. (18). We construct this model as follows: if  $\theta_{\text{fid}}$  is the vector of the parameter values determining our fiducial cosmology, given by the parameters in the first row of Table I, and the associated best-fit HOD parameters from Eq. (18), we model the parameter dependence using the following expansion:

<sup>16</sup>The validity of this assumption was recently tested in the weak lensing application of Ref. [61], in which the probability distribution of the WST coefficients was found to be closer to Gaussian than that of the bispectrum. Also see Ref. [114] for an in-depth analysis of the validity of this assumption for various statistics.



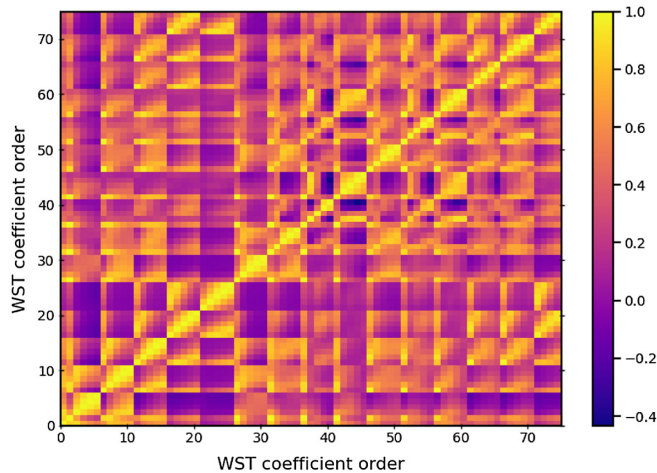


FIG. 3. Correlation matrix of all 76 coefficients for the WST evaluated at the fiducial cosmology. The WST coefficients populate the data vector in order of increasing values of the  $j_1$  and  $l_1$  indices, with the  $l_1$  index varied faster, as in Fig. 2.

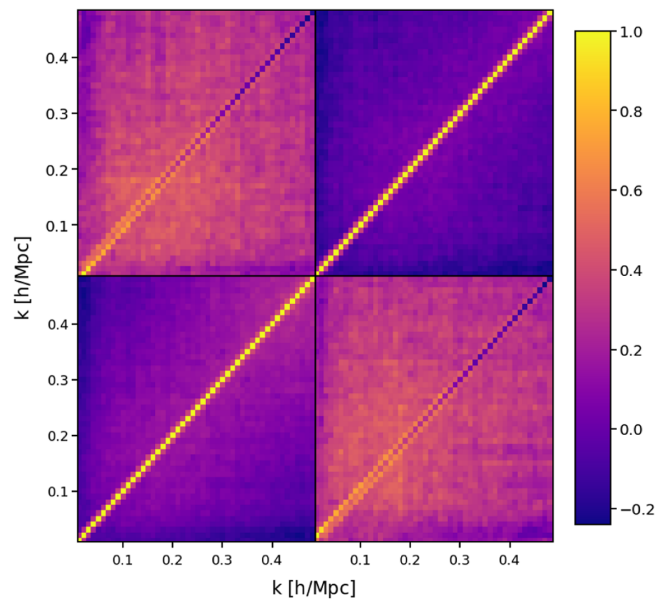


FIG. 4. Correlation matrix of the galaxy power spectrum multipoles,  $l = \{0, 2\}$ , evaluated from the 2048 realizations of the PATCHY mocks for the fiducial cosmology. In the  $2 \times 2$  blocks, from bottom to top and from left to right, we visualize the auto and cross correlations of  $\hat{P}_0$  and  $\hat{P}_2$ , respectively.

$$\mathbf{X}_r(\theta) = \mathbf{X}_r(\theta_{\text{fid}}) + (\theta - \theta_{\text{fid}})\nabla_{\theta}\mathbf{X}, \quad (35)$$

where the gradient  $\nabla_{\theta}\mathbf{X}$  in Eq. (35) is straightforwardly determined using the derivatives constructed from the derivative grid cosmologies of Table I, as we explained in Sec. III B 1. We have carefully checked and confirmed that the combination of the vector dimensionality, derivative step size and number of HOD realizations used is sufficient for the derivatives to be well-converged, for all

parameters and for both estimators (see Appendix C for details). It should be clarified, at this point, that being a first order expansion, Eq. (35) needs to be evaluated using a fiducial cosmology  $\theta_{\text{fid}}$  sufficiently close to the true one, such that the derivative approximation only breaks down far away from the (true) maximum of the likelihood and the correct cosmology is recovered after a likelihood analysis of the data. Since our chosen  $\theta_{\text{fid}}$  corresponds to the mean values of the *Planck* 2018 [83]  $\Lambda$ CDM cosmology, and as it can be also inferred by the very good agreement between the fiducial theory prediction and the one from the data shown in Figs. 1 and 2, this requirement is satisfied in our analysis. We do, however, point out that since the Taylor expansion (35) will inevitably break down far away from the fiducial part of the parameter space, our model will fail to capture any potential non-Gaussianities in the likelihood, for which a full model (e.g., an emulator) for  $\mathbf{X}_r(\theta)$  would be necessary. We add that a very similar expansion was also recently used in the Dark Energy Survey (DES) year-1 data re-analysis of Ref. [116], which employed a hybrid combination of perturbation theory and ABACUSSUMMIT simulations in order to model the lensing power spectrum needed for their analysis.<sup>17</sup> We leave the construction of an actual WST emulator for future work. Finally, we clarify that, even though a wide variety of approaches exist for a more accurate modeling of the cosmological dependence of the galaxy power spectrum or correlation function (e.g., perturbation theory [20,76–78,109,110] or emulators [90,103,117–119]), we use the model from Eq. (35) for the power spectrum, as well, in order to guarantee a fair comparison against the performance of the WST.

In order to perform a posterior analysis, we sample the likelihood from Eq. (35) using the Markov Chain Monte Carlo (MCMC) sampler *emcee* [120],<sup>18</sup> choosing 100 walkers, 500 “burn-in” steps and 100,000 steps for our main runs. We use flat unrestricted priors (with the walkers initialized in the range 0–1.05) for all parameters, with the exception of  $\omega_b$ , for which we also consider the case of a Gaussian prior determined from big bang nucleosynthesis (BBN) [83] and the measurement of helium and deuterium primordial abundances [121–123]:

$$\omega_b = 0.02268 \pm 0.00038, \quad (36)$$

a choice commonly adopted in analyses of BOSS data [76–78]. Finally, to check the convergence of our chains, we monitor the mean integrated autocorrelation time and make sure its value is at least 2 orders of magnitude lower

<sup>17</sup>We note that Ref. [116] was able to correct the errors caused by such a Taylor expansion, using the HALOFIT model for the matter power spectrum. Given that no such possibility is available for the cosmological dependence of the WST coefficients, a correction of this kind was not possible in our case.

<sup>18</sup>Publicly available in <https://emcee.readthedocs.io/en/stable/>.



than the total number of steps used, following the procedure laid out by Ref. [120]. We similarly also monitor the mean value of the acceptance fraction and make sure it always falls within the reasonable range of values of 0.3–0.5.

#### IV. RESULTS

Before presenting our results from the likelihood analysis explained in Sec. III E, we briefly summarize the procedure we follow in order to determine the posterior of the Hubble constant,  $h$ , as a derived parameter. As stated in Sec. III B 1, in the ABACUSSUMMIT simulations the value of the angular size of the sound horizon at last scattering,  $\theta_*$ , is kept fixed to the corresponding mean value derived from measurements by the *Planck* satellite,  $100\theta_* = 1.041533$ . This implies that we can compute  $h$  for each point in our chains and derive its posterior, which we will also discuss below as the fifth cosmological parameter determined from our analysis.

We now discuss the results of our analysis of the BOSS CMASS data using the WST and the power spectrum multipoles, starting with the case in which the BBN prior (36) was imposed on the value of  $\omega_b$ . In Fig. 5, we show the resulting 2-dimensional posterior probability distribution function of the three other cosmological parameters explicitly varied in the likelihood (32), together with the derived Hubble constant  $h$ , all of which have been marginalized

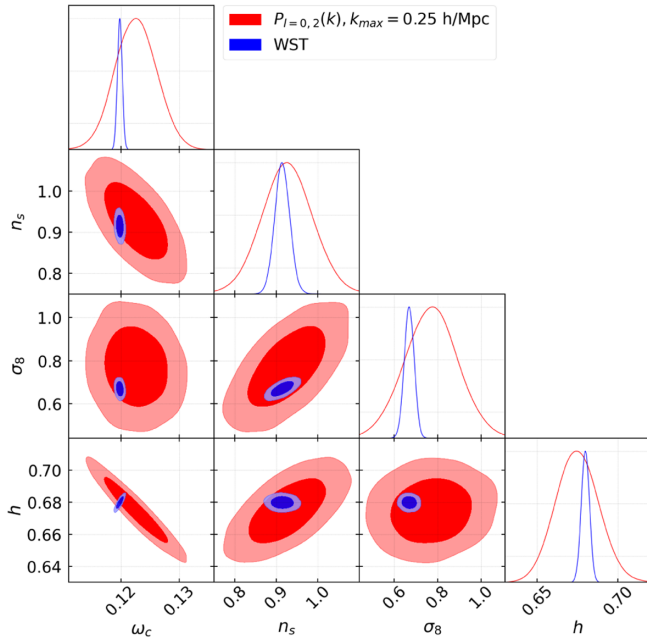


FIG. 5. Constraints on the cosmological parameters obtained from the combined monopole and quadrupole of the galaxy power spectrum evaluated up to  $k_{\max} = 0.25$  Mpc/h (red contours), as well as from the WST coefficients defined in Sec. III D (blue contours). The results shown above were obtained after imposing a BBN Gaussian prior on the value of  $\omega_b = 0.02268 \pm 0.00038$ .

over the seven nuisance parameters of the HOD model. In addition, the mean and  $1\sigma$  error values obtained from the two estimators for all cosmological parameters are listed in Table II, while the full corner plot is presented in Appendix D.

We notice the relative consistency between the corresponding mean values for the parameters extracted from the two estimators, the differences of which never exceed the respective  $1\sigma$  values from the power spectrum. Furthermore, all values are broadly consistent with the ones found from recent re-analyses of BOSS data [20,76–78,109,110], a fact that confirms the robustness of the WST as a tool to be used for cosmological analyses. More importantly, in addition to being able to serve as a reliable clustering statistic to infer cosmological parameters from the LSS, the WST is found to deliver significant improvements in the inferred  $1\sigma$  errors for all cosmological parameters, in the range  $3 - 8\times$  tighter. This finding demonstrates the potential carried in the use of the WST as a way to access the non-Gaussian information encoded in the LSS data (as suggested in our previous work [52]), and thus subsequently improve the errors obtained on cosmological parameters.

In Fig. 6 we show a case with no priors on any of the cosmological parameters (the results are also summarized in Table II). We observe a similar trend as in the previous case, with the WST once again outperforming the regular power spectrum with respect to the obtained  $1\sigma$  errors, by a factor of 3–5. The inferred mean values of the 5 cosmological parameters are once again consistent, within  $1\sigma$ , between the two statistics, despite the fact that the totally unrestricted priors led to a relatively lower value of  $\omega_b$  (relative to the BBN prior), and subsequently of  $h$  (through the fixed  $\theta_*$ ).

We also briefly comment on the results obtained from a likelihood analysis using the vector of WST coefficients up to first order, only. In this case, the inferred  $1\sigma$  errors from the WST are improved compared to the corresponding power spectrum results by a factor of 1.1 to 2.0.

A fair comparison with previous analyses of the BOSS data [20,76–78,109,110,124,125] is difficult given several

TABLE II. Mean values and 68% confidence intervals for all cosmological parameters resulting from the posterior analysis of the power spectrum multipoles and the WST coefficients in the case of a BBN prior applied on the value of  $\omega_b$  (left half), and the case of unrestricted priors (right half). All results are presented in the format  $\text{mean}_{-1\sigma}^{+1\sigma}$ , after marginalization over all HOD parameters.

	BBN prior on $\omega_b$		Unrestricted priors	
	P(k)	WST	P(k)	WST
$\omega_b$	$0.02268_{-0.00036}^{+0.00036}$	$0.02225_{-0.00034}^{+0.00034}$	$0.0217_{-0.0043}^{+0.0043}$	$0.0184_{-0.0011}^{+0.0011}$
$\omega_c$	$0.1225_{-0.0037}^{+0.0037}$	$0.120_{-0.00041}^{+0.00041}$	$0.1217_{-0.0058}^{+0.0058}$	$0.1154_{-0.0012}^{+0.0012}$
$n_s$	$0.927_{-0.063}^{+0.063}$	$0.914_{-0.018}^{+0.018}$	$0.921_{-0.049}^{+0.057}$	$0.931_{-0.018}^{+0.018}$
$\sigma_8$	$0.77_{-0.13}^{+0.13}$	$0.67_{-0.023}^{+0.023}$	$0.762_{-0.094}^{+0.11}$	$0.691_{-0.023}^{+0.023}$
$h$	$0.675_{-0.014}^{+0.014}$	$0.68_{-0.0025}^{+0.0025}$	$0.668_{-0.024}^{+0.024}$	$0.653_{-0.0074}^{+0.0074}$

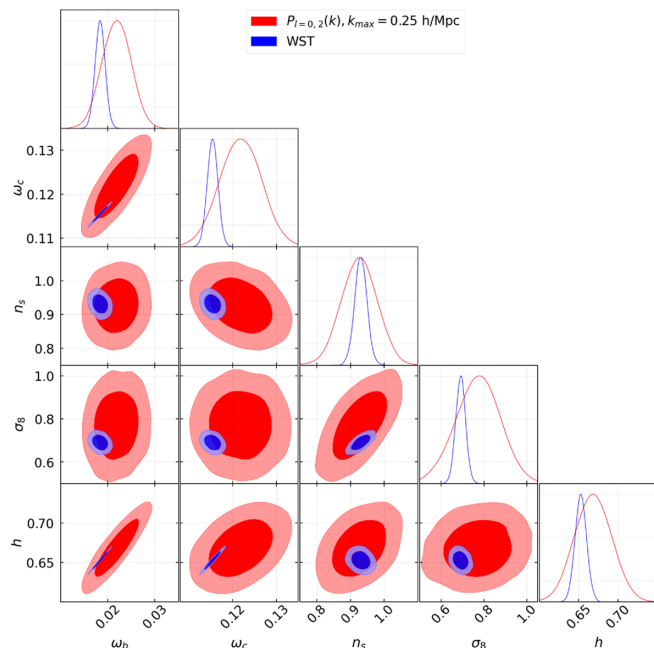


FIG. 6. Same as in Fig. 5, but using a flat and uninformative prior on  $\omega_b$ .

differences between those works and ours, the major ones being the approximate model we employed from Eq. (35) and the fact that the Hubble constant was not explicitly varied in our analysis, but was rather derived from a fixed  $\theta_*$ .<sup>19</sup> This is also the case when attempting to compare against recent BOSS analyses using emulators of the redshift-space power spectrum [126] or the correlation function [90,119], with the additional difference that the correlation function analyses focused on much smaller scales than the ones we worked with. It should be also noted that the  $1\sigma$  errors reported above are purely statistical, since we did not attempt to quantify how the various systematics and approximations adopted by our analysis would affect the final results.

Furthermore, we clarify that in our WST analysis we have purely worked with unreconstructed density fields. In the context of traditional BOSS analyses, reconstruction algorithms [127] have been shown to enable a more precise determination of the baryon acoustic oscillation (BAO) peak position, the complementary information of which can further improve the constraints obtained using the full shape of the power spectrum [78,109] and also the bispectrum [20]. Whether a WST analysis applied to the reconstructed density field can improve the constraints extracted on cosmological parameters is an interesting question, that we plan to explore in the future. Putting these differences aside, we do highlight that the relative  $1\sigma$  error from the

<sup>19</sup>We also worked with a subset of the CMASS sample, rather than with the full CMASS and LOWZ samples, due to the limitations imposed by the HOD procedure.

power spectrum obtained on the Hubble constant is found to be equal to 2%, when using the BBN prior on  $\omega_b$ , a value that is similar to the one recently found by Refs. [20,76,78,109].

Finally, we finish this section by commenting on the fact that the mean value of  $\sigma_8$  obtained from our WST application is in very good agreement with results from recent BOSS reanalyses, which are also found to be in tension with the corresponding *Planck* value, in particular for the case of an unrestricted prior on  $n_s$  [20]. Imposing a *Planck* prior on  $n_s$  somewhat raises the recovered mean value of  $\sigma_8$  [20,109], but is not large enough to completely alleviate the tension. Furthermore, cross-correlating BOSS clustering data with CMB lensing measured by *Planck* has been recently found to further lower the inferred value of  $\sigma_8$  [124].

## V. CONCLUSIONS

In this work, we present the first application of the wavelet scattering transform on actual galaxy observations, through a WST analysis of the BOSS DR12 CMASS dataset.

Building upon our previous LSS application to 3D matter overdensity fields [52], we lay out the detailed methodology to capture additional layers of realism that are necessary to analyze galaxy observations obtained from a spectroscopic survey, such as BOSS. After capturing the effects of redshift-space anisotropy, nontrivial survey geometry, the shortcomings of the dataset through a set of systematic weights and the Alcock-Paczynski effect, we show how to transform a galaxy sample from redshift-space sky coordinates into the weighted Feldman-Kaiser-Peacock (FKP) field, which serve as the input of a WST scattering network. The resulting WST coefficients can then be treated as a well-defined basis that reflects the clustering properties of the observed sample, which we use as the main object of our BOSS analysis.

In order to model the cosmological dependence of the WST coefficients we use the state-of-the-art suite of ABACUSSUMMIT simulations [79]. These span the cosmological parameter space around the *Planck* 2018  $\Lambda$ CDM cosmology [83], and have been HOD-tuned to match small-scale redshift-space correlation function of the BOSS CMASS sample in the redshift range  $0.46 < z < 0.60$ . For the evaluation of the WST covariance matrix, which is also necessary in addition to the model, we employ the publicly available MULTIDARK-PATCHY mocks [87,91]. We take all necessary steps to ensure that our mock theory predictions satisfy the same level of realism as the observations we compare them against, and also evaluate the multipoles of the anisotropic galaxy power spectrum, which we use as a benchmark to assess the performance of the WST.

We then use our model to perform a likelihood analysis of the CMASS observations with the WST coefficients and the power spectrum multipoles. We obtain the posterior probability distributions of the 4 target cosmological parameters,  $\{\omega_b, \omega_c, n_s, \sigma_8\}$ , as well as the Hubble parameter,  $h$ , derived from the fixed value  $\theta_*$ , all of which were

marginalized over the seven nuisance parameters of the HOD model. The analysis reveals a substantial improvement in the values of the  $1\sigma$  errors predicted by the WST, which are tighter than the corresponding ones from the regular power spectrum by a factor of 3–8, when a BBN prior is applied on the value of  $\omega_b$ , and by a factor in the range 3–5 in the case of flat and uninformative priors for all of the parameters. At the same time, the inferred mean values of all cosmological parameters by the WST (as well as the power spectrum) are always found to be broadly consistent with the ones found by recent re-analyses of BOSS data [20,76–78,109,110,124], demonstrating, overall, that the WST can be reliably used as a powerful estimator in modern analyses of LSS data.<sup>20</sup>

We should take note, at this point, of certain limitations of our current analysis, that we plan to tackle in a follow-up work. First, the Taylor expansion from Eq. (35), that we used to emulate the parameter dependence of our model vector in the likelihood [Eq. (32)] is inevitably expected to break down in parameter regions far away from the fiducial cosmology, a fact that essentially prevents us from capturing any substantial non-Gaussianities present in the likelihood. In order to do this, a full model for the nonlinear dependence of the WST coefficients as a function of the cosmological parameters will need to be developed, similar to the training procedure of an emulator for a given clustering statistic. We envision that well-established emulation techniques (such as the one presented in Ref. [90] for the correlation function, or emulators at the field level [128]) could be straightforwardly expanded to enable a full WST application, such as the one we performed in this work. Second, the Hubble parameter was not explicitly varied in our analysis, but rather obtained as a derived parameter through the fixed angular scale  $\theta_*$  (to the value measured by the *Planck* satellite [83]) in the ABACUSSUMMIT simulations. This limitation can be easily overcome by using a different set of mocks, in which  $h$  is explicitly varied. Furthermore, the current set of ABACUS mocks we used did not account for the effect of light-cone evolution of the galaxy clustering within the survey volume,<sup>21</sup> an effect that can be rather easily overcome using the next generation of ABACUS mocks already in production. Lastly, the weighting scheme (13) and (14), which we adopted to correct for the data systematics, was designed for the power spectrum case, rather than for the WST. Even though this option is expected to capture these effects to some extent, a correction scheme tailored to the WST estimator would be preferred. After all of the above improvements are implemented, a fair comparison against the results obtained by state-of-the-art reanalyses of BOSS

data using perturbation theory ([20,76–78,109,110,124]) or emulators ([90,103,117–119,126]) will finally be possible, a step that we reserve for future work.

Our BOSS analysis hints at a wide range of exciting future applications of the WST in the context of LSS cosmology. In our previous work [52], we showed that the WST coefficients are particularly sensitive to the properties of massive neutrinos, thanks to their innate ability to capture clustering information beyond the traditional 2-point function, combined with their ability to trace the properties of voids. Even though we did not include neutrinos in our current analysis, since they were not varied in the ABACUSSUMMIT simulations, they can be easily incorporated using a future set of mocks that captures their effect, which could then potentially allow us to obtain powerful constraints on the sum of the neutrino masses. This is also the case for a variety of other  $\Lambda$ CDM extensions, such as theories for modified gravity or dynamical dark energy. We also emphasize here that the procedure we laid out is very flexible, and can be straightforwardly applied to any future set of spectroscopic galaxy observations, given an associated set of systematic weights and mock catalogs, both of which are commonly produced to support analyses using traditional estimators. For example, and subject to the additional improvements discussed above, our framework can be easily adjusted for a future application to spectroscopic observations by DESI.

We note that despite their impressive performance in the context of an LSS analysis, the wavelets we used to implement the scattering network from Eq. (6) were proposed in the context of a 3D molecular chemistry application [65,66]. One could envision developing wavelets optimized for a cosmological application, which can further improve the benefits of a WST analysis. For example, equivariant wavelets [72] can find a natural application in the case of fields with a particular directional dependence, as is the galaxy overdensity observed in redshift space, that we have used as the input field in the current work.

In addition, we comment on the fact that the second order WST used in this analysis encodes information from correlation functions up to 4th order. It would be interesting, as a result, to compare a BOSS analysis using WST and the 4-point correlation function [21] in future work.

Through this first application of the wavelet scattering transform on actual galaxy observations, we demonstrate that this technique can serve as a promising tool for current and future applications of cosmological parameter inference.

## ACKNOWLEDGMENTS

C.D. is partially supported by NSF Grant No. AST-1813694. This work is supported by the National Science Foundation under Cooperative Agreement No. PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions). We would like to thank Sihan Yuan for kindly sharing the first generation of ABACUSSUMMIT mocks with us, for providing help with

<sup>20</sup>We note that this result relies on the adoption of a reliable and accurate HOD model.

<sup>21</sup>This effect is already captured, however, by the covariance from the PATCHY mocks.



the use of ABACUSHOD and for discussions over the course of this work. We are grateful to Boryana Hadzhiyska, Lehman H. Garrison and Daniel J. Eisenstein for valuable discussions and information on the suite of ABACUSSUMMIT simulations and the associated galaxy mocks. We would also like to thank Arthur Tsang for detailed comments on our draft. Furthermore, we would like to thank Florian Beutler and Hector Gil Marin for useful discussions on the BOSS data and also Sihao Cheng and Tom Abel for sharing insights on the wavelet scattering transform. The massive production of all MultiDark-Patchy mocks for the BOSS Final Data Release has been performed at the BSC Marenostrum supercomputer, the Hydra cluster at the Instituto de Fisica Teorica UAM/CSIC, and NERSC at the Lawrence Berkeley National Laboratory. That work acknowledges support from the Spanish MICINN's Consolider-Ingenio 2010 Programme under Grant No. MultiDark CSD2009-00064, MINECO Centro de Excelencia Severo Ochoa Programme under grant No. SEV-2012-0249, and grant No. AYA2014-60641-C2-1-P. The MultiDark-Patchy mocks was an effort led from the IFT UAM-CSIC by F. Prada's group (C.-H. Chuang, S. Rodriguez-Torres and C. Scoccola) in collaboration with C. Zhao (Tsinghua U.), F.-S. Kitaura (AIP), A. Klypin (NMSU), G. Yepes (UAM), and the BOSS galaxy clustering working group.

### APPENDIX A: WST FOR MASKED DENSITY FIELDS

In the original WST implementation (3) in KYMATIO, the input field is assumed to be a periodic 3D cube, such as, for example, the output of an N-body simulation. Given that in this application, however, we work with data that occupy the nontrivial survey geometry of BOSS, we need to make the necessary modifications. In particular, and in direct analogy to the power spectrum case, we start by embedding the masked density fields (23) (both from the data and the cut-sky mocks) into 3D cubes using `nbodykit`. These 3D grids, which contain both the actual volume of the survey and also the part of the cube that lies outside of the BOSS mask, are fed as input into KYMATIO. We then need to make sure that the fundamental WST evaluations, wavelet convolution, modulus and averaging, include only the contributions from regions within the mask. In practice, we modify KYMATIO such that regions of the 3D grid that lie outside the mask do not contribute to the convolutions in Eq. (3), in any order. This is straightforward to implement with minimal modifications, given the exact knowledge of the survey binary mask. Likewise, the regions outside the mask are always zeroed out and do not receive any contributions from the density field through the wavelet convolutions. Finally, we take the modulus and average over the parts of the field that lie within the survey footprint, in order to get the WST coefficients from a

masked input field. This procedure is analogous to the corresponding evaluation of the power spectrum monopole for an input masked field.

In addition to occupying an irregular survey geometry, we also note that the density fields we work with in this application are anisotropic, due to the effects of RSD, with the survey line-of-sight lying along the radial direction in a spherical coordinate system. Given that the basis of solid harmonic wavelets we adopted was designed for an isotropic input field, without treating any direction as special, the current WST configuration might not fully leverage all the information encoded in the RSD field (similar to evaluating only the monopole of the power spectrum, that averages over all directions). Such a shortcoming is indeed possible to overcome, for example with the directional-dependent equivariant wavelets of Ref. [72]. We defer this study for future work.

### APPENDIX B: NUMERICAL CONVERGENCE

Given that the covariance matrix of the WST coefficients is evaluated from simulations, we need to make sure that the number of realizations used is sufficient to guarantee the numerical convergence of the results. We show in Fig. 7 the  $1-\sigma$  errors on the cosmological parameters as a function of the number of PATCHY mock realizations used to evaluate the covariance matrix. We find that the change (relative to the results obtained from the full suite of  $N_{\text{cov}} = 2048$  realizations) is smaller than 1.8% for all parameters, when using  $N_{\text{cov}} \geq 1800$  realizations. This confirms the numerical convergence of the WST covariance.

### APPENDIX C: DATA COMPRESSION

In addition to the covariance matrix convergence discussed in the previous section, we also confirm the numerical stability of the derivatives entering the Taylor expansion in Eq. (35). After testing how the  $1-\sigma$  errors obtained on the cosmological parameters change as a function of the number

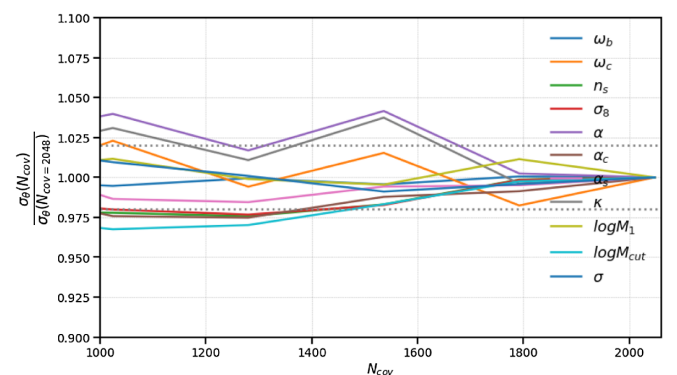


FIG. 7. The  $1-\sigma$  errors on the cosmological parameters,  $\sigma_\theta$ , plotted as a function of the number of PATCHY mock realizations,  $N_{\text{cov}}$ , used to evaluate the WST covariance matrix. The y axis is normalized with respect to the  $1-\sigma$  errors obtained when using the total number of available  $N_{\text{cov}} = 2048$  realizations.



of realizations used to evaluate the derivatives, we found variations of at most 6% when using half of the total realizations. Furthermore, we note again that the simulations used to construct the derivatives were run with phase-matched initial conditions. This fact, in combination also with the much higher volume and resolution of ABACUS compared to other existing simulations, is expected to modulate the noise due to cosmic variance, at least to some extent. Given, however, the relatively low number of 20 HOD realizations available for the evaluation of these derivatives, we also test the numerical stability by repeating our analysis using a compressed data vector.

In particular, for a likelihood  $\mathcal{L}(\theta|\mathbf{d})$ , the quantity [129]

$$t = \nabla_{\theta} \mathcal{L}(\theta|\mathbf{d}), \quad (\text{C1})$$

corresponds to a compression from the original data vector of dimensionality  $N_d$  down to one with dimensions equal to the number of parameters,  $n$ . In the case of a Gaussian likelihood with a data covariance independent of the cosmological parameters, Eq. (C1) is further simplified to

$$t = \nabla_{\theta} \mathbf{X}^T C^{-1} [\mathbf{X}_d - \mathbf{X}_t(\theta_{\text{fid}})], \quad (\text{C2})$$

which represents a linear and lossless compression that preserves the Fisher matrix of the original estimator [130,131]. The Gaussian likelihood of the compressed statistic is

$$\log \mathcal{L}_t(\theta|\mathbf{d}) = -\frac{1}{2} [t - t(\theta)]^T C_t^{-1} [t - t(\theta)] + \text{const.}, \quad (\text{C3})$$

where

$$t(\theta) = \nabla_{\theta} \mathbf{X}^T C^{-1} [\mathbf{X}_t(\theta) - \mathbf{X}_t(\theta_{\text{fid}})], \quad (\text{C4})$$

and with  $C_t$  the covariance matrix of  $t$ . In the particular case of the WST, the original data vector of  $N_d = 76$  coefficients is compressed down to  $n = 11$  numbers. Reductions of this kind greatly reduce the challenges associated with parameter inference from high-dimensional data vectors and have been utilized in bispectrum applications [132–134]. More importantly for our case, the compression (C1) leads to a statistic that is less sensitive to numerical noise, being a weighted average of the original data points. Indeed, such a compression has been recently used to accelerate the convergence of Fisher forecasts, the numerical stability of which is notoriously challenging in the case of noisy derivatives [135–137].

We compress the WST and the power spectrum multipole vectors using Eqs. (C2)–(C4) and then repeat the parameter inference application of the analysis section, sampling from the likelihood (C3). In Fig. 8, we compare the marginalized 2-dimensional posteriors obtained for the cosmological parameters using the compressed WST, against the results obtained from the original, uncompressed, WST analysis in the case of a BBN prior. The very small differences between these two sets of results serve as additional confirmation of the robustness of the

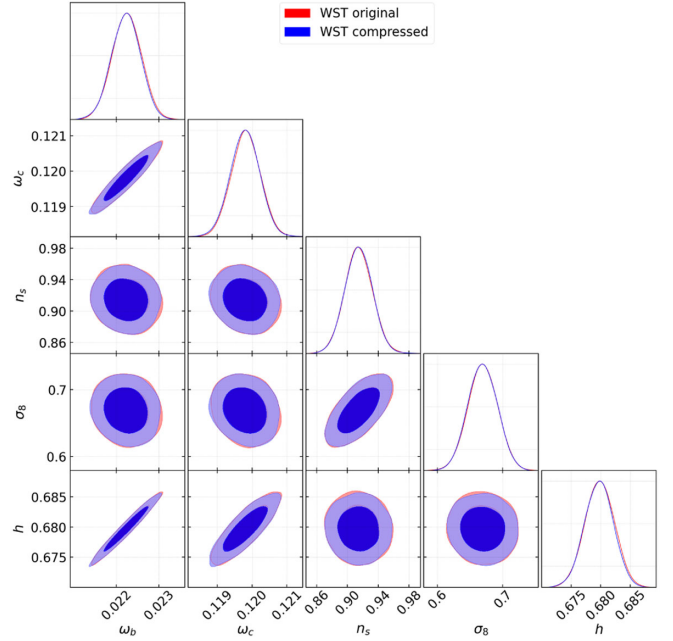


FIG. 8. Constraints on the cosmological parameters obtained from the original WST coefficients defined in Sec. III D (red contours), as well as from the compressed version of the WST data vector from Eqs. (C2)–(C4) (blue contours). The results shown above were obtained after imposing a BBN Gaussian prior on the value of  $\omega_b = 0.02268 \pm 0.00038$ .

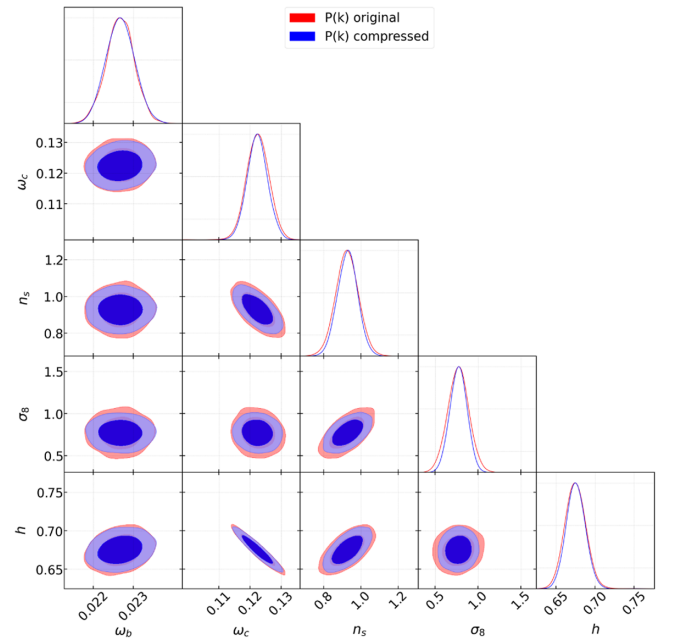


FIG. 9. Constraints on the cosmological parameters obtained from the galaxy power spectrum multipoles (red contours), as well as from the compressed version of the power spectrum data vector from Eqs. (C2)–(C4) (blue contours). The results shown above were obtained after imposing a BBN Gaussian prior on the value of  $\omega_b = 0.02268 \pm 0.00038$ .

main analysis. Finally, in Fig. 9 we show the same comparison for the multipoles of the galaxy power spectrum, reaching a similar conclusion.

**APPENDIX D: FULL PARAMETER SPACE FOR P(k) AND WST**

For completeness, in this appendix we show the full corner plots from our likelihood analysis, including the

marginalized posteriors of the seven parameters of our HOD model, that we treated as nuisance parameters. In particular, in Fig. 10 we show the full corner plot of the analysis with a BBN prior on the value of  $\omega_b$ , while Fig. 11 illustrates the results for the case of unrestricted priors on all parameters. We see that the 12 parameters shown are consistent (within  $1\sigma$ ) for the mean values recovered using the WST coefficients and the power spectrum multipoles.

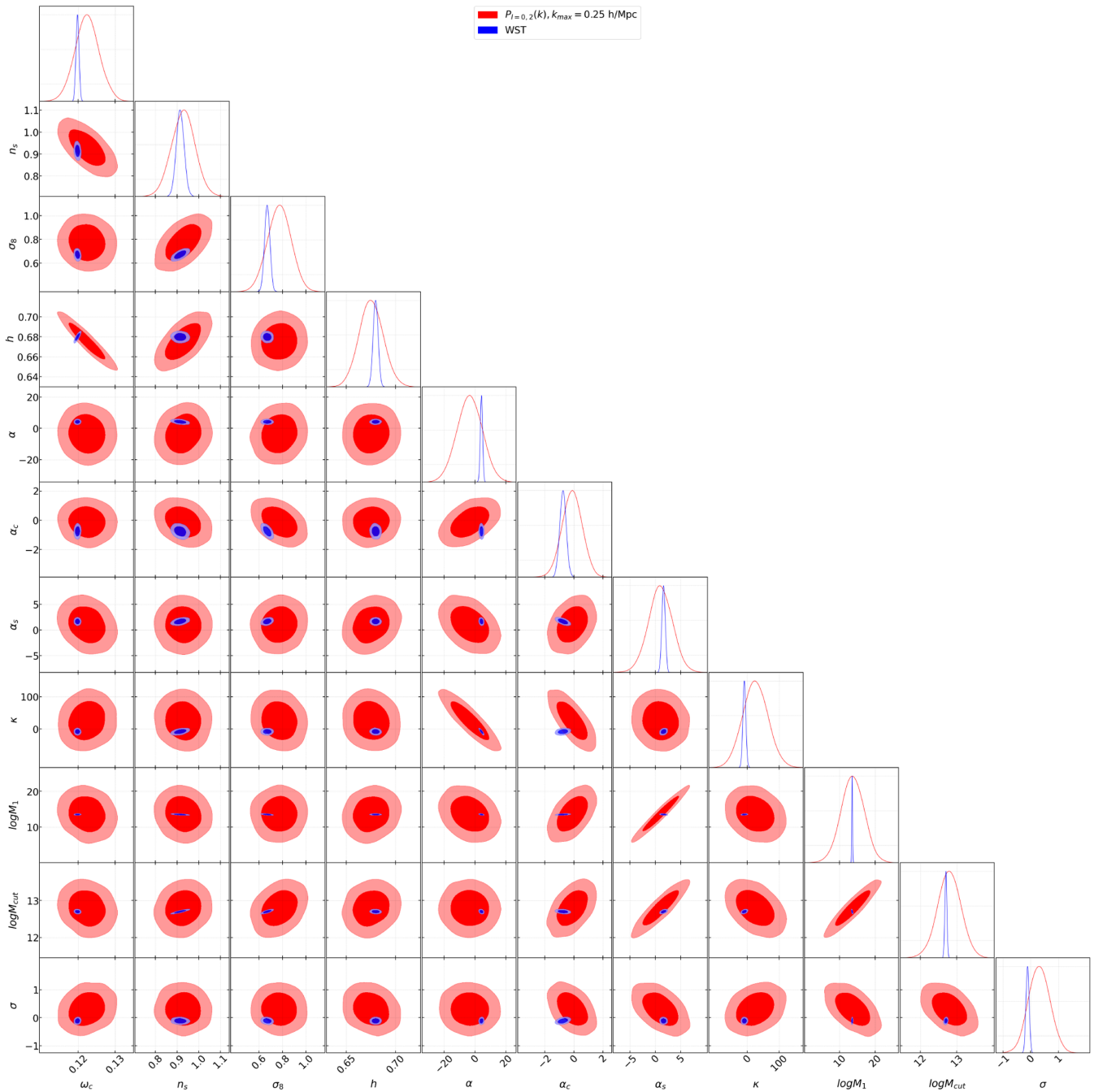


FIG. 10. Same as in Fig. 5, but now including the 7 parameters of the HOD model.  $M_{\text{cut}}$  and  $M_1$  are expressed in units of  $M_{\odot}/h$ .

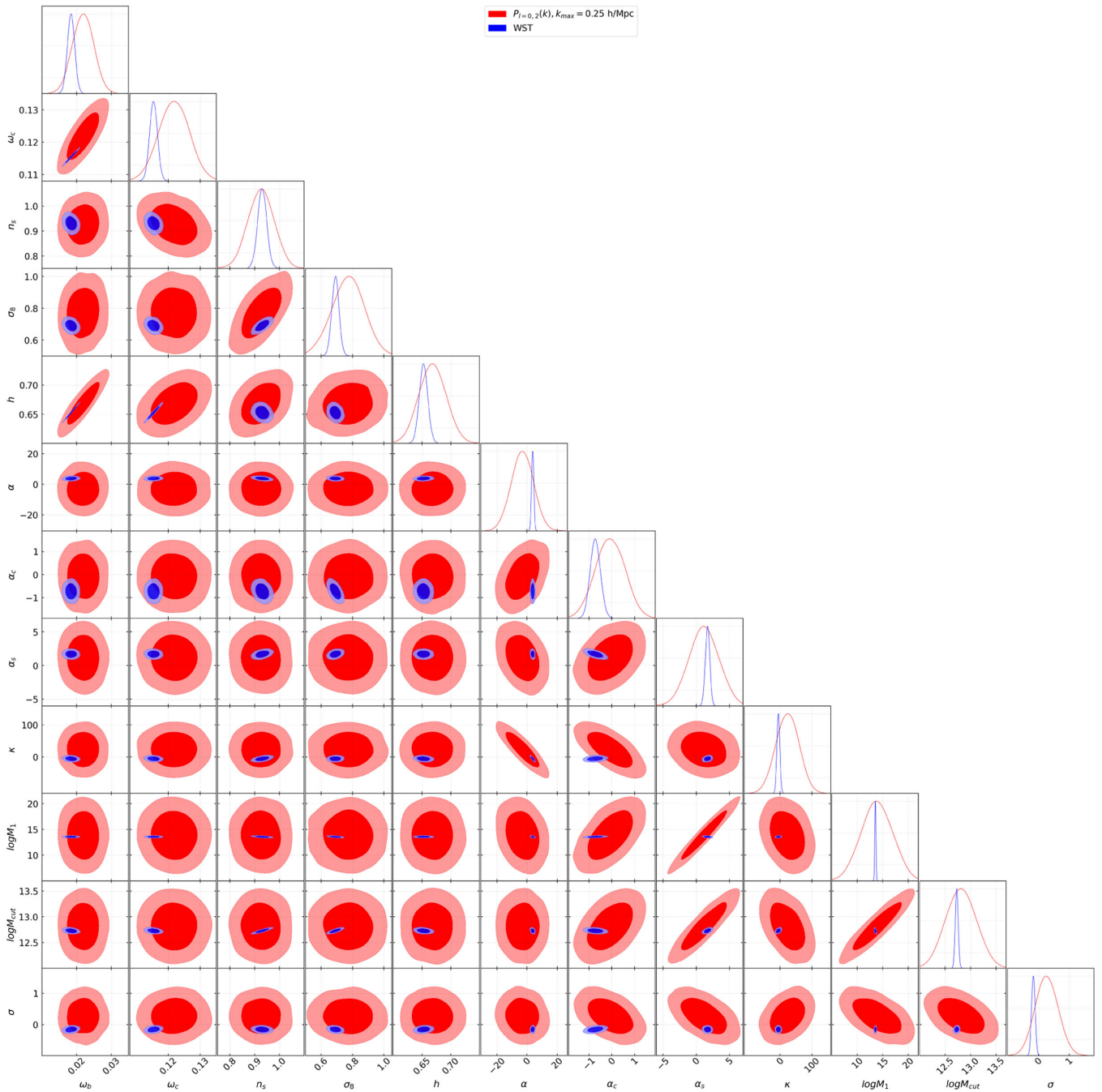


FIG. 11. Same as in Fig. 6, but now including the seven parameters of the HOD model.  $M_{\text{cut}}$  and  $M_1$  are expressed in units of  $M_{\odot}/h$ .

- 
- [1] E. J. Copeland, M. Sami, and S. Tsujikawa, Dynamics of dark energy, *Int. J. Mod. Phys. D* **15**, 1753 (2006).
  - [2] M. Ishak, Testing general relativity in cosmology, *Living Rev. Relativity* **22**, 1 (2019).
  - [3] P. G. Ferreira, Cosmological tests of gravity, *Annu. Rev. Astron. Astrophys.* **57**, 335 (2019).
  - [4] S. Alam *et al.*, Testing the theory of gravity with DESI: Estimators, predictions and simulation requirements, *J. Cosmol. Astropart. Phys.* **11** (2021) 050.
  - [5] A. Drlica-Wagner *et al.* (LSST Dark Matter Group), Probing the fundamental nature of dark matter with the large synoptic survey telescope, [arXiv:1902.01055](https://arxiv.org/abs/1902.01055).

- [6] J. Lesgourgues and S. Pastor, Massive neutrinos and cosmology, *Phys. Rep.* **429**, 307 (2006).
- [7] C. Dvorkin *et al.*, Neutrino mass from cosmology: Probing physics beyond the standard model, [arXiv:1903.03689](https://arxiv.org/abs/1903.03689).
- [8] X. Chen, C. Dvorkin, Z. Huang, M. H. Namjoo, and L. Verde, The future of primordial features with large-scale structure surveys, *J. Cosmol. Astropart. Phys.* **11** (2016) 014.
- [9] M. Levi *et al.* (DESI Collaboration), The DESI Experiment, a whitepaper for Snowmass 2013, [arXiv:1308.0847](https://arxiv.org/abs/1308.0847).
- [10] P. A. Abell *et al.* (LSST Science Collaborations, LSST Project), LSST Science Book, Version 2.0, [arXiv:0912.0201](https://arxiv.org/abs/0912.0201).
- [11] A. Abate *et al.* (LSST Dark Energy Science), Large synoptic survey telescope: Dark energy science collaboration, [arXiv:1211.0310](https://arxiv.org/abs/1211.0310).
- [12] R. Laureijs *et al.* (EUCLID Collaboration), Euclid definition study report, [arXiv:1110.3193](https://arxiv.org/abs/1110.3193).
- [13] D. Spergel, N. Gehrels, J. Breckinridge, M. Donahue, A. Dressler *et al.*, Wide-field infrared survey telescope—astrophysics focused telescope assets WFIRST-AFTA final report, [arXiv:1305.5422](https://arxiv.org/abs/1305.5422).
- [14] J. Carron, Information Escaping the Correlation Hierarchy of the Convergence Field in the Study of Cosmological Parameters, *Phys. Rev. Lett.* **108**, 071301 (2012).
- [15] H. Gil-Marín, J. Noreña, L. Verde, W. J. Percival, C. Wagner, M. Manera, and D. P. Schneider, The power spectrum and bispectrum of SDSS DR11 BOSS galaxies—I. Bias and gravity, *Mon. Not. R. Astron. Soc.* **451**, 539 (2015).
- [16] H. Gil-Marín, W. J. Percival, L. Verde, J. R. Brownstein, C.-H. Chuang, F.-S. Kitaura, S. A. Rodríguez-Torres, and M. D. Olmstead, The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the power spectrum and bispectrum of the DR12 BOSS galaxies, *Mon. Not. R. Astron. Soc.* **465**, 1757 (2017).
- [17] F. Bernardeau, S. Colombi, E. Gaztañaga, and R. Scoccimarro, Large-scale structure of the universe and cosmological perturbation theory, *Phys. Rep.* **367**, 1 (2002).
- [18] C. Hahn, F. Villaescusa-Navarro, E. Castorina, and R. Scoccimarro, Constraining  $m$  with the bispectrum. Part I. Breaking parameter degeneracies, *J. Cosmol. Astropart. Phys.* **03** (2020) 040.
- [19] C. Hahn and F. Villaescusa-Navarro, Constraining  $m$  with the bispectrum. part II. the information content of the galaxy bispectrum monopole, *J. Cosmol. Astropart. Phys.* **04** (2021) 029.
- [20] O. H. E. Philcox and M. M. Ivanov, Boss dr12 full-shape cosmology:  $\Lambda$ CDM constraints from the large-scale galaxy power spectrum and bispectrum monopole, *Phys. Rev. D* **105**, 043517 (2022).
- [21] O. H. E. Philcox, J. Hou, and Z. Slepian, A first detection of the connected 4-point correlation function of galaxies using the BOSS CMASS sample, [arXiv:2108.01670](https://arxiv.org/abs/2108.01670).
- [22] V. Desjacques, D. Jeong, and F. Schmidt, Large-scale galaxy bias, *Phys. Rep.* **733**, 1 (2018), large-scale galaxy bias.
- [23] A. J. S. Hamilton, Linear redshift distortions: A review, in *The Evolving Universe: Selected Topics on Large-Scale Structure and on the Properties of Galaxies*, edited by D. Hamilton (Springer, Netherlands, Dordrecht, 1998), pp. 185–275.
- [24] H. A. Feldman, N. Kaiser, and J. A. Peacock, Power-spectrum analysis of three-dimensional redshift surveys, *Astrophys. J.* **426**, 23 (1994).
- [25] M. C. Neyrinck, I. Szapudi, and A. S. Szalay, Rejuvenating the matter power spectrum: Restoring information with a logarithmic density mapping, *Astrophys. J.* **698**, L90 (2009).
- [26] F. Simpson, J. B. James, A. F. Heavens, and C. Heymans, Clipping the Cosmos: The Bias and Bispectrum of Large Scale Structure, *Phys. Rev. Lett.* **107**, 271301 (2011).
- [27] M. White, A marked correlation function for constraining modified gravity models, *J. Cosmol. Astropart. Phys.* **11** (2016) 057.
- [28] G. Valogiannis and R. Bean, Beyond  $\delta$ : Tailoring marked statistics to reveal modified gravity, *Phys. Rev. D* **97**, 023535 (2018).
- [29] E. Massara, F. Villaescusa-Navarro, S. Ho, N. Dalal, and D. N. Spergel, Using the Marked Power Spectrum to Detect the Signature of Neutrinos in Large-Scale Structure, *Phys. Rev. Lett.* **126**, 011301 (2021).
- [30] A. Pisani *et al.*, Cosmic voids: A novel probe to shed light on our Universe, [arXiv:1903.05161](https://arxiv.org/abs/1903.05161).
- [31] E. Massara, F. Villaescusa-Navarro, M. Viel, and P. Sutter, Voids in massive neutrino cosmologies, *J. Cosmol. Astropart. Phys.* **11** (2015) 018.
- [32] C. D. Kreisch, A. Pisani, C. Carbone, J. Liu, A. J. Hawken, E. Massara, D. N. Spergel, and B. D. Wandelt, Massive neutrinos leave fingerprints on cosmic voids, *Mon. Not. R. Astron. Soc.* **488**, 4413 (2019).
- [33] Y.-C. Cai, N. Padilla, and B. Li, Testing gravity using cosmic voids, *Mon. Not. R. Astron. Soc.* **451**, 1036 (2015).
- [34] N. Hamaus, P. Sutter, G. Lavaux, and B. D. Wandelt, Probing cosmology and gravity with redshift-space distortions around voids, *J. Cosmol. Astropart. Phys.* **11** (2015) 036.
- [35] C. D. Kreisch, A. Pisani, F. Villaescusa-Navarro, D. N. Spergel, B. D. Wandelt, N. Hamaus, and A. E. Bayer, The GIGANTES dataset: Precision cosmology from voids in the machine learning era, *Astrophys. J.* **935**, 100 (2022).
- [36] T. Bonnaire, N. Aghanim, J. Kuruville, and A. Decelle, Cosmology with cosmic web environments I. Real-space power spectra, *Astron. Astrophys.* **661**, A146 (2022).
- [37] M. Schmittfull, T. Baldauf, and U. c. v. Seljak, Near optimal bispectrum estimators for large-scale structure, *Phys. Rev. D* **91**, 043530 (2015).
- [38] A. M. Dizgah, H. Lee, M. Schmittfull, and C. Dvorkin, Capturing non-Gaussianity of the large-scale structure with weighted skew-spectra, *J. Cosmol. Astropart. Phys.* **04** (2020) 011.
- [39] P. Chakraborty, S.-F. Chen, and C. Dvorkin, Skewing the CMB  $\times$  LSS: A fast method for bispectrum analysis, *J. Cosmol. Astropart. Phys.* **07** (2022) 038.
- [40] J. Schmalzing, S. Gottlöber, A. A. Klypin, and A. V. Kravtsov, Quantifying the evolution of higher order clustering, *Mon. Not. R. Astron. Soc.* **309**, 1007 (1999).
- [41] G. Pratten and D. Munshi, Non-Gaussianity in large-scale structure and Minkowski functionals, *Mon. Not. R. Astron. Soc.* **423**, 3209 (2012).



- [42] S. Codis, C. Pichon, D. Pogosyan, F. Bernardeau, and T. Matsubara, Non-Gaussian Minkowski functionals and extrema counts in redshift space, *Mon. Not. R. Astron. Soc.* **435**, 531 (2013).
- [43] W. Fang, B. Li, and G.-B. Zhao, New Probe of Departures from General Relativity using Minkowski Functionals, *Phys. Rev. Lett.* **118**, 181301 (2017).
- [44] A. Banerjee and T. Abel, Nearest neighbour distributions: New statistical measures for cosmological clustering, *Mon. Not. R. Astron. Soc.* **500**, 5479 (2020).
- [45] K. Naidoo, E. Massara, and O. Lahav, Cosmology and neutrino mass with the minimum spanning tree, *Mon. Not. R. Astron. Soc.* **513**, 3596 (2022).
- [46] C. Uhlemann, O. Friedrich, F. Villaescusa-Navarro, A. Banerjee, and S. Codis, Fisher for complements: Extracting cosmology and neutrino mass from the counts-in-cells PDF, *Mon. Not. R. Astron. Soc.* **495**, 4006 (2020).
- [47] J. Bruna and S. Mallat, Invariant scattering convolution networks, *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1872 (2013).
- [48] A. Gupta, J. M. Z. Matilla, D. Hsu, and Z. Haiman, Non-Gaussian information from weak lensing data via deep learning, *Phys. Rev. D* **97**, 103515 (2018).
- [49] F. Villaescusa-Navarro, D. Anglés-Alcázar, S. Genel, D. N. Spergel, R. S. Somerville, R. Dave, A. Pillepich, L. Hernquist, D. Nelson, P. Torrey, D. Narayanan, Y. Li, O. Philcox, V. L. Torre, A. M. Delgado, S. Ho *et al.*, The CAMELS project: Cosmology and astrophysics with machine-learning simulations, *Astrophys. J.* **915**, 71 (2021).
- [50] L. A. Perez, S. Genel, F. Villaescusa-Navarro, R. S. Somerville, A. Gabrielpillai, D. Anglés-Alcázar, B. D. Wandelt, and L. Y. A. Yung, Constraining cosmology with machine learning and galaxy clustering: The CAMELS-SAM suite, [arXiv:2204.02408](https://arxiv.org/abs/2204.02408).
- [51] C. Dvorkin *et al.*, Machine learning and cosmology, [arXiv:2203.08056](https://arxiv.org/abs/2203.08056).
- [52] G. Valogiannis and C. Dvorkin, Towards an optimal estimation of cosmological parameters with the wavelet scattering transform, *Phys. Rev. D* **105**, 103534 (2022).
- [53] S. Mallat, Group invariant scattering, *Commun. Pure Appl. Math.* **65**, 1331 (2012).
- [54] L. Sifre and S. Mallat, Rotation, scaling and deformation invariant scattering for texture discrimination, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Portland, Oregon, 2013).
- [55] J. Bruna, S. Mallat, E. Bacry, and J.-F. Muzy, Intermittent process analysis with scattering moments, *Ann. Stat.* **43**, 323 (2015).
- [56] J. Andén and S. Mallat, Deep scattering spectrum, *IEEE Trans. Signal Process.* **62**, 4114 (2014).
- [57] E. Allys, F. Levrier, S. Zhang, C. Colling, B. Regaldo-Saint Blancard, F. Boulanger, P. Hennebelle, and S. Mallat, The RWST, a comprehensive statistical description of the non-Gaussian structures in the ISM, *Astron. Astrophys.* **629**, A115 (2019).
- [58] A. K. Saydjari, S. K. N. Portillo, Z. Slepian, S. Kahraman, B. Burkhart, and D. P. Finkbeiner, Classification of magnetohydrodynamic simulations using wavelet scattering transforms, *Astrophys. J.* **910**, 122 (2021).
- [59] B. Regaldo-Saint Blancard, F. Levrier, E. Allys, E. Bellomi, and F. Boulanger, Statistical description of dust polarized emission from the diffuse interstellar medium—a RWST approach, *Astron. Astrophys.* **642**, A217 (2020).
- [60] S. Cheng, Y.-S. Ting, B. Ménard, and J. Bruna, A new approach to observational cosmology using the scattering transform, *Mon. Not. R. Astron. Soc.* **499**, 5902 (2020).
- [61] S. Cheng and B. Menard, Weak lensing scattering transform: Dark energy and neutrino mass sensitivity, *Mon. Not. R. Astron. Soc.* **507**, 1012 (2021).
- [62] E. Allys, T. Marchand, J.-F. Cardoso, F. Villaescusa-Navarro, S. Ho, and S. Mallat, New interpretable statistics for large-scale structure analysis and generation, *Phys. Rev. D* **102**, 103506 (2020).
- [63] B. Greig, Y.-S. Ting, and A. A. Kaurov, Exploring the cosmic 21-cm signal from the epoch of reionisation using the wavelet scattering transform, *Mon. Not. R. Astron. Soc.* **513**, 1719 (2022).
- [64] M. Eickenberg, E. Allys, A. Moradinezhad Dizgah, P. Lemos, E. Massara, M. Abidi, C. Hahn, S. Hassan, B. Regaldo-Saint Blancard, S. Ho, S. Mallat, J. Anden, and F. Villaescusa-Navarro, Wavelet moments for cosmological parameter estimation, [arXiv:2204.07646](https://arxiv.org/abs/2204.07646).
- [65] M. Eickenberg, G. Exarchakis, M. Hirn, and S. Mallat, Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3d electronic densities, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17* (Curran Associates Inc., Red Hook, NY, USA, 2017), p. 6543–6552.
- [66] M. Eickenberg, G. Exarchakis, M. Hirn, S. Mallat, and L. Thiry, Solid harmonic wavelet scattering for predictions of molecule properties, *J. Chem. Phys.* **148**, 241732 (2018).
- [67] S. Cheng and B. Menard, How to quantify fields or textures? A guide to the scattering transform, [arXiv:2112.01288](https://arxiv.org/abs/2112.01288).
- [68] F. Villaescusa-Navarro, C. Hahn, E. Massara, A. Banerjee, A. M. Delgado, D. K. Ramanah, T. Charnock, E. Giusarma, Y. Li, E. Allys, A. Brochard, C. Uhlemann, C.-T. Chiang, S. He, A. Pisani, A. Obuljen *et al.*, The quiote simulations, *Astrophys. J. Suppl. Ser.* **250**, 2 (2020).
- [69] S. Alam, F. D. Albareti, C. A. Prieto, F. Anders, S. F. Anderson, T. Anderton, B. H. Andrews, E. Armengaud, É. Aubourg, S. Bailey, S. Basu, J. E. Bautista, R. L. Beaton, T. C. Beers, C. F. Bender, A. A. Berlind *et al.*, The eleventh and twelfth data releases of the Sloan digital sky survey: Final data from SDSS-III, *Astrophys. J. Suppl. Ser.* **219**, 12 (2015).
- [70] D. J. Eisenstein, D. H. Weinberg, E. Agol, H. Aihara, C. A. Prieto, S. F. Anderson, J. A. Arns, É. Aubourg, S. Bailey, E. Balbinot, R. Barkhouser, T. C. Beers, A. A. Berlind, S. J. Bickerton, D. Bizyaev, M. R. Blanton *et al.*, SDSS-III: Massive spectroscopic surveys of the distant universe, the milky way, and extra-solar planetary systems, *Astron. J.* **142**, 72 (2011).
- [71] K. S. Dawson, D. J. Schlegel, C. P. Ahn, S. F. Anderson, É. Aubourg, S. Bailey, R. H. Barkhouser, J. E. Bautista, A. Beifiori, A. A. Berlind, V. Bhardwaj, D. Bizyaev, C. H. Blake, M. R. Blanton, M. Blomqvist, A. S. Bolton *et al.*,

- The baryon oscillation spectroscopic survey of SDSS-III, *Astron. J.* **145**, 10 (2013).
- [72] A. K. Saydjari and D. P. Finkbeiner, Equivariant wavelets: Fast rotation and translation invariant wavelet scattering transforms, [arXiv:2104.11244](https://arxiv.org/abs/2104.11244).
- [73] H. Gil-Marín, W. J. Percival, J. R. Brownstein, C.-H. Chuang, J. N. Grieb, S. Ho, F.-S. Kitaura, C. Maraston, F. Prada, S. Rodríguez-Torres, A. J. Ross, L. Samushia, D. J. Schlegel, D. Thomas, J. L. Tinker, and G.-B. Zhao, The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: RSD measurement from the LOS-dependent power spectrum of DR12 BOSS galaxies, *Mon. Not. R. Astron. Soc.* **460**, 4188 (2016).
- [74] F. Beutler, H.-J. Seo, S. Saito, C.-H. Chuang, A. J. Cuesta, D. J. Eisenstein, H. Gil-Marín, J. N. Grieb, N. Hand, F.-S. Kitaura, C. Modi, R. C. Nichol, M. D. Olmstead, W. J. Percival, F. Prada, A. G. Sánchez *et al.*, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Anisotropic galaxy clustering in Fourier space, *Mon. Not. R. Astron. Soc.* **466**, 2242 (2017).
- [75] C. Alcock and B. Paczynski, An evolution free test for non-zero cosmological constant, *Nature (London)* **281**, 358 (1979).
- [76] M. M. Ivanov, M. Simonović, and M. Zaldarriaga, Cosmological parameters from the BOSS galaxy power spectrum, *J. Cosmol. Astropart. Phys.* **05** (2020) 042.
- [77] G. d’Amico, J. Gleyzes, N. Kokron, K. Markovic, L. Senatore, P. Zhang, F. Beutler, and H. Gil-Marín, The cosmological analysis of the SDSS/BOSS data from the effective field theory of large-scale structure, *J. Cosmol. Astropart. Phys.* **05** (2020) 005.
- [78] O. H. Philcox, M. M. Ivanov, M. Simonović, and M. Zaldarriaga, Combining full-shape and BAO analyses of galaxy power spectra: A 1.6% CMB-independent constraint on  $h_0$ , *J. Cosmol. Astropart. Phys.* **05** (2020) 032.
- [79] N. A. Maksimova, L. H. Garrison, D. J. Eisenstein, B. Hadzhiyska, S. Bose, and T. P. Satterthwaite, AbacusSummit: A massive set of high-accuracy, high-resolution N-body simulations, *Mon. Not. R. Astron. Soc.* **508**, 4017 (2021).
- [80] L. H. Garrison, D. J. Eisenstein, and P. A. Pinto, A high-fidelity realization of the Euclid code comparison N-body simulation with Abacus, *Mon. Not. R. Astron. Soc.* **485**, 3370 (2019).
- [81] L. H. Garrison, D. J. Eisenstein, D. Ferrer, N. A. Maksimova, and P. A. Pinto, The abacus cosmological N-body code, *Mon. Not. R. Astron. Soc.* **508**, 575 (2021).
- [82] B. Hadzhiyska, D. Eisenstein, S. Bose, L. H. Garrison, and N. Maksimova, compaso: A new halo finder for competitive assignment to spherical overdensities, *Mon. Not. R. Astron. Soc.* **509**, 501 (2021).
- [83] N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J.-P. Bernard, M. Bersanelli, P. Bielewicz *et al.* (Planck Collaboration), Planck 2018 results—vi. cosmological parameters, *Astron. Astrophys.* **641**, A6 (2020).
- [84] S. Yuan, L. H. Garrison, B. Hadzhiyska, S. Bose, and D. J. Eisenstein, AbacusHOD: A highly efficient extended multitracer HOD framework and its application to BOSS and eBOSS data, *Mon. Not. R. Astron. Soc.* **510**, 3301 (2022).
- [85] Z. Zheng, A. L. Coil, and I. Zehavi, Galaxy evolution from halo occupation distribution modeling of DEEP2 and SDSS galaxy clustering, *Astrophys. J.* **667**, 760 (2007).
- [86] B. Reid, S. Ho, N. Padmanabhan, W. J. Percival, J. Tinker, R. Tojeiro, M. White, D. J. Eisenstein, C. Maraston, A. J. Ross, A. G. Sánchez, D. Schlegel, E. Sheldon, M. A. Strauss, D. Thomas, D. Wake *et al.*, SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 12: Galaxy target selection and large-scale structure catalogs, *Mon. Not. R. Astron. Soc.* **455**, 1553 (2016).
- [87] S. A. Rodríguez-Torres, C.-H. Chuang, F. Prada, H. Guo, A. Klypin, P. Behroozi, C. H. Hahn, J. Comparat, G. Yepes, A. D. Montero-Dorta, J. R. Brownstein, C. Maraston, C. K. McBride, J. Tinker, S. Gottlöber, G. Favole *et al.*, The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Modelling the clustering and halo occupation distribution of BOSS CMASS galaxies in the Final Data Release, *Mon. Not. R. Astron. Soc.* **460**, 1173 (2016).
- [88] S. Yuan, B. Hadzhiyska, S. Bose, and D. J. Eisenstein, Illustrating galaxy-halo connection in the DESI era with IllustrisTNG, *Mon. Not. R. Astron. Soc.* **512**, 5793 (2022).
- [89] S. Yuan, B. Hadzhiyska, S. Bose, D. J. Eisenstein, and H. Guo, Evidence for galaxy assembly bias in BOSS CMASS redshift-space galaxy correlation function, *Mon. Not. R. Astron. Soc.* **502**, 3582 (2021).
- [90] S. Yuan, L. H. Garrison, D. J. Eisenstein, and R. H. Wechsler, Stringent  $\sigma_8$  constraints from small-scale galaxy clustering using a hybrid MCMC + emulator framework, *Mon. Not. R. Astron. Soc.* **515**, 871 (2022).
- [91] F.-S. Kitaura, S. Rodríguez-Torres, C.-H. Chuang, C. Zhao, F. Prada, H. Gil-Marín, H. Guo, G. Yepes, A. Klypin, C. G. Scóccola, J. Tinker, C. McBride, B. Reid, A. G. Sánchez, S. Salazar-Albornoz, J. N. Grieb *et al.*, The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Mock galaxy catalogs for the BOSS Final Data Release, *Mon. Not. R. Astron. Soc.* **456**, 4156 (2016).
- [92] A. Klypin, G. Yepes, S. Gottlöber, F. Prada, and S. Heß, MultiDark simulations: The story of dark matter halo concentrations and density profiles, *Mon. Not. R. Astron. Soc.* **457**, 4340 (2016).
- [93] V. Springel, S. D. M. White, A. Jenkins, C. S. Frenk, N. Yoshida, L. Gao, J. Navarro, R. Thacker, D. Croton, J. Helly, J. A. Peacock, S. Cole, P. Thomas, H. Couchman, A. Evrard, J. Colberg *et al.*, Simulations of the formation, evolution and clustering of galaxies and quasars, *Nature (London)* **435**, 629 (2005).
- [94] A. Klypin and J. Holtzman, Particle mesh code for cosmological simulations, [arXiv:astro-ph/9712217](https://arxiv.org/abs/astro-ph/9712217).
- [95] A. V. Kravtsov, A. A. Berlind, R. H. Wechsler, A. A. Klypin, S. Gottlöber, B. Allgood, and J. R. Primack, The dark side of the halo occupation distribution, *Astrophys. J.* **609**, 35 (2004).
- [96] D. Kodwani, D. Alonso, and P. Ferreira, The effect on cosmological parameter estimation of a parameter dependent covariance matrix, *Open J. Astrophys.* **2** (2019).

- [97] J. Carron, On the assumption of Gaussianity for cosmological two-point statistics and parameter dependent covariance matrices, *Astron. Astrophys.* **551**, A88 (2013).
- [98] K. Yamamoto, M. Nakamichi, A. Kamino, B. A. Bassett, and H. Nishioka, A measurement of the quadrupole power spectrum in the clustering of the 2dF QSO survey, *Publ. Astron. Soc. Jpn.* **58**, 93 (2006).
- [99] D. Bianchi, H. Gil-Marín, R. Ruggeri, and W. J. Percival, Measuring line-of-sight-dependent Fourier-space clustering using FFTs, *Mon. Not. R. Astron. Soc.* **453**, L11 (2015).
- [100] R. Scoccimarro, Fast estimators for redshift-space clustering, *Phys. Rev. D* **92**, 083532 (2015).
- [101] N. Hand, Y. Li, Z. Slepian, and U. Seljak, An optimal FFT-based anisotropic power spectrum estimator, *J. Cosmol. Astropart. Phys.* **07** (2017) 002.
- [102] R. Hockney and J. Eastwood, *Computer Simulation Using Particles*, Advanced Book Program (McGraw-Hill International Book Company, New York, 1981).
- [103] Y. Kobayashi, T. Nishimichi, M. Takada, R. Takahashi, and K. Osato, Accurate emulator for the redshift-space power spectrum of dark matter halos and its application to galaxy power spectrum, *Phys. Rev. D* **102**, 063504 (2020).
- [104] M. J. Wilson, J. A. Peacock, A. N. Taylor, and S. de la Torre, Rapid modelling of the redshift-space power spectrum multipoles for a masked density field, *Mon. Not. R. Astron. Soc.* **464**, 3121 (2017).
- [105] F. Beutler, S. Saito, H.-J. Seo, J. Brinkmann, K. S. Dawson, D. J. Eisenstein, A. Font-Ribera, S. Ho, C. K. McBride, F. Montesano, W. J. Percival, A. J. Ross, N. P. Ross, L. Samushia, D. J. Schlegel, A. G. Sánchez *et al.*, The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Testing gravity with redshift space distortions using the power spectrum multipoles, *Mon. Not. R. Astron. Soc.* **443**, 1065 (2014).
- [106] A. J. S. Hamilton, Uncorrelated modes of the non-linear power spectrum, *Mon. Not. R. Astron. Soc.* **312**, 257 (2000).
- [107] F. Beutler and P. McDonald, Unified galaxy power spectrum measurements from 6dfgs, BOSS, and eBOSS, *J. Cosmol. Astropart. Phys.* **11** (2021) 031.
- [108] J. U. Lange, A. P. Hearin, A. Leauthaud, F. C. van den Bosch, H. Guo, and J. DeRose, Five per cent measurements of the growth rate from simulation-based modelling of redshift-space clustering in BOSS LOWZ, *Mon. Not. R. Astron. Soc.* **509**, 1779 (2021).
- [109] S.-F. Chen, Z. Vlah, and M. White, A new analysis of galaxy 2-point functions in the BOSS survey, including full-shape information and post-reconstruction BAO, *J. Cosmol. Astropart. Phys.* **02** (2022) 008.
- [110] P. Zhang, G. D'Amico, L. Senatore, C. Zhao, and Y. Cai, BOSS correlation function analysis from the effective field theory of large-scale structure, *J. Cosmol. Astropart. Phys.* **02** (2022) 036.
- [111] M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky, J. Bruna, V. LOSTANLEN, M. J. Hirn, E. Oyallon, S. Zhang, C. Cella *et al.*, Kymatio: Scattering transforms in Python, [arXiv:1812.11214](https://arxiv.org/abs/1812.11214).
- [112] F. G. Mohammad, F. Villaescusa-Navarro, S. Genel, D. Angles-Alcazar, and M. Vogelsberger, Inpainting hydrodynamical maps with deep learning, [arXiv:2109.07070](https://arxiv.org/abs/2109.07070).
- [113] M. White, J. L. Tinker, and C. K. McBride, Mock galaxy catalogues using the quick particle mesh method, *Mon. Not. R. Astron. Soc.* **437**, 2594 (2014).
- [114] C. F. Park, E. Allys, F. Villaescusa-Navarro, and D. P. Finkbeiner, Quantification of high dimensional non-Gaussianities and its implication to Fisher analysis in cosmology, [arXiv:2204.05435](https://arxiv.org/abs/2204.05435).
- [115] J. Hartlap, P. Simon, and P. Schneider, Why your model parameter confidences might be too optimistic. unbiased estimation of the inverse covariance matrix, *Astron. Astrophys.* **464**, 399 (2007).
- [116] B. Hadzhiyska, C. García-García, D. Alonso, A. Nicola, and A. Slosar, Hefty enhancement of cosmological constraints from the DES y1 data using a hybrid effective field theory approach to galaxy bias, *J. Cosmol. Astropart. Phys.* **09** (2021) 020.
- [117] T. Nishimichi, M. Takada, R. Takahashi, K. Osato, M. Shirasaki, T. Oogi, H. Miyatake, M. Oguri, R. Murata, Y. Kobayashi, and N. Yoshida, Dark quest. I. Fast and accurate emulation of halo clustering statistics and its application to galaxy clustering, *Astrophys. J.* **884**, 29 (2019).
- [118] H. Miyatake, Y. Kobayashi, M. Takada, T. Nishimichi, M. Shirasaki, S. Sugiyama, R. Takahashi, K. Osato, S. More, and Y. Park, Cosmological inference from emulator based halo model I: Validation tests with HSC and SDSS mock catalogs, *Phys. Rev. D* **106**, 083519 (2022).
- [119] Z. Zhai, J. L. Tinker, A. Banerjee, J. DeRose, H. Guo, Y.-Y. Mao, S. McLaughlin, K. Storey-Fisher, and R. H. Wechsler, The Aemulus Project V: Cosmological constraint from small-scale clustering of BOSS galaxies, [arXiv:2203.08999](https://arxiv.org/abs/2203.08999).
- [120] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: The MCMC Hammer, *Publ. Astron. Soc. Pac.* **125**, 306 (2013).
- [121] E. Aver, K. A. Olive, and E. D. Skillman, The effects of the I lambda 10830 on helium abundance determinations, *J. Cosmol. Astropart. Phys.* **07** (2015) 011.
- [122] R. J. Cooke, M. Pettini, and C. C. Steidel, One percent determination of the primordial deuterium abundance, *Astrophys. J.* **855**, 102 (2018).
- [123] N. Schöneberg, J. Lesgourgues, and D. C. Hooper, The bao + bbn take on the Hubble tension, *J. Cosmol. Astropart. Phys.* **10** (2019) 029.
- [124] S.-F. Chen, M. White, J. DeRose, and N. Kokron, Cosmological analysis of three-dimensional BOSS galaxy clustering and Planck CMB lensing cross correlations via Lagrangian perturbation theory, *J. Cosmol. Astropart. Phys.* **07** (2022) 041.
- [125] S. Brieden, H. Gil-Marín, and L. Verde, Model-agnostic interpretation of 10 billion years of cosmic evolution traced by BOSS and eBOSS data, *J. Cosmol. Astropart. Phys.* **08** (2022) 024.
- [126] Y. Kobayashi, T. Nishimichi, M. Takada, and H. Miyatake, Full-shape cosmology analysis of SDSS-III BOSS galaxy power spectrum using emulator-based halo model: A 5% determination of  $\sigma_8$ , *Phys. Rev. D* **105**, 083517 (2022).



- [127] D. J. Eisenstein, H.-J. Seo, E. Sirko, and D. N. Spergel, Improving cosmological distance measurements by reconstruction of the baryon acoustic peak, *Astrophys. J.* **664**, 675 (2007).
- [128] N. Kaushal, F. Villaescusa-Navarro, E. Giusarma, Y. Li, C. Hawry, and M. Reyes, NECOLA: Towards a universal field-level cosmological emulator, *Astrophys. J.* **930**, 115 (2022).
- [129] J. Alsing and B. Wandelt, Generalized massive optimal data compression, *Mon. Not. R. Astron. Soc.* **476**, L60 (2018).
- [130] M. Tegmark, A. N. Taylor, and A. F. Heavens, Karhunen-Loeve eigenvalue problems in cosmology: How should we tackle large data sets?, *Astrophys. J.* **480**, 22 (1997).
- [131] A. F. Heavens, R. Jimenez, and O. Lahav, Massive lossless data compression and multiple parameter estimation from galaxy spectra, *Mon. Not. R. Astron. Soc.* **317**, 965 (2000).
- [132] D. Gualdi, H. Gil-Marín, R. L. Schuhmann, M. Manera, B. Joachimi, and O. Lahav, Enhancing BOSS bispectrum cosmological constraints with maximal compression, *Mon. Not. R. Astron. Soc.* **484**, 3713 (2019).
- [133] O. H. E. Philcox, M. M. Ivanov, M. Zaldarriaga, M. Simonovic, and M. Schmittfull, Fewer mocks and less noise: Reducing the dimensionality of cosmological observables with subspace projections, *Phys. Rev. D* **103**, 043508 (2021).
- [134] J. Byun and E. Krause, Modal compression of the redshift-space galaxy bispectrum, *Mon. Not. R. Astron. Soc. stac2313* (2022).
- [135] W. R. Coulton, F. Villaescusa-Navarro, D. Jamieson, M. Baldi, G. Jung, D. Karagiannis, M. Liguori, L. Verde, and B. D. Wandelt, Quijote PNG: The information content of the halo power spectrum and bispectrum, [arXiv:2206.15450](https://arxiv.org/abs/2206.15450).
- [136] W. R. Coulton, F. Villaescusa-Navarro, D. Jamieson, M. Baldi, G. Jung, D. Karagiannis, M. Liguori, L. Verde, and B. D. Wandelt, Quijote-PNG: Simulations of primordial non-Gaussianity and the information content of the matter field power spectrum and bispectrum, [arXiv:2206.01619](https://arxiv.org/abs/2206.01619).
- [137] G. Jung, D. Karagiannis, M. Liguori, M. Baldi, W. R. Coulton, D. Jamieson, L. Verde, F. Villaescusa-Navarro, and B. D. Wandelt, Quijote-PNG: Quasi-maximum likelihood estimation of Primordial Non-Gaussianity in the non-linear dark matter density field, [arXiv:2206.01624](https://arxiv.org/abs/2206.01624).