Precise cosmological constraints from BOSS galaxy clustering with a simulation-based emulator of the wavelet scattering transform

Georgios Valogiannis,[1,2,*] Sihan Yuan,[3,4,†] and Cora Dvorkin[1,‡]

[1]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*Department of Astronomy and Astrophysics, University of Chicago, Chicago, Illinois 60637, USA*
[3]*Kavli Institute for Particle Astrophysics and Cosmology, 452 Lomita Mall, Stanford University, Stanford, California 94305, USA*
[4]*SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA*

(Received 27 October 2023; accepted 11 April 2024; published 2 May 2024)

We perform a reanalysis of the BOSS CMASS DR12 galaxy dataset using a simulation-based emulator for the wavelet scattering transform (WST) coefficients. Moving beyond our previous works, which laid the foundation for the first galaxy clustering application of this estimator, we construct a neural net-based emulator for the cosmological dependence of the WST coefficients and the 2-point correlation function multipoles, trained from the state-of-the-art suite of AbacusSummit simulations combined with a flexible halo occupation distribution (HOD) galaxy model. In order to confirm the accuracy of our pipeline, we subject it to a series of thorough internal and external mock parameter recovery tests, before applying it to reanalyze the CMASS observations in the redshift range $0.46 < z < 0.57$. We find that a joint WST + 2-point correlation function likelihood analysis allows us to obtain marginalized 1σ errors on the ΛCDM parameters that are tighter by a factor of 2.5–6, compared to the 2-point correlation function, and by a factor of 1.4–2.5 compared to the WST-only results. This corresponds to a competitive 0.9%, 2.3% and 1% level of determination for parameters ω_c, σ_8 & n_s, respectively, and also to a 0.7% and 2.5% constraint on derived parameters h and $f(z)\sigma_8(z)$, in agreement with the *Planck* 2018 results. Our results reaffirm the constraining power of the WST and highlight the exciting prospect of employing higher-order statistics in order to fully exploit the power of upcoming stage-IV spectroscopic observations.

DOI: 10.1103/PhysRevD.109.103503

I. INTRODUCTION

The advent of precision cosmology, with a large collection of surveys including the Dark Energy Spectroscopic Instrument (DESI) [1,2], the *Vera C. Rubin* Observatory Legacy Survey of Space and Time (LSST) [3,4], *Euclid* [5], and the *Nancy Grace Roman* Space Telescope [6], that will accurately probe the 3-dimensional (3D) large-scale structure (LSS) of the universe, promises to dramatically change our fundamental understanding of the cosmos. Among the wealth of valuable information offered by cosmological observations of this kind, lies the opportunity to tackle major open questions in modern physics, such as the source of the accelerated expansion of the universe at late times [7], the nature of dark matter [8], the large-scale properties of gravity [9–11], the properties of massive neutrinos [12,13], as well as the physics of the primordial universe and other light relics [14–16].

The probability distribution that describes the observed large-scale structure of the universe at late times is known to deviate from the familiar Gaussian form characterizing the primordial density field. The nonlinear process of gravitational instability, responsible for the formation of the 3D cosmic web, imparts a non-Gaussian distribution in the observed large-scale structure of the universe. As a consequence, the standard compression achieved by the 2-point correlation function of density fluctuations fails to capture all available information encoded in the clustered field [17]. Even though working with the 2-point function statistics has sufficed in traditional applications of cosmological parameter inference up until recently, such an approach will be inadequate if the potential of the upcoming generation of cosmological surveys is to be fully exploited. Accurately modeling structure formation down to the nonlinear regime in principle requires the inclusion of higher-order moments as a part of the traditional parameter inference, a line of research that is currently very actively pursued [18–26]. Nevertheless, the requirements associated with handling n-point correlation functions, both in terms of the necessary computational cost of evaluation, but also due to the relatively large dimensionality of the final data

*gvalogiannis@g.harvard.edu
†sihany@stanford.edu
‡cdvorkin@g.harvard.edu

vector, quickly render such an approach intractable when going to higher n. Even when these challenges can be tempered using different kinds of techniques, the total information encoded in a non-Guassian field has been shown to escape the entire correlation hierarchy, with the magnitude of loss getting progressively more pronounced with increasing degree of non-Gaussianity [17].

The obstacles mentioned above motivate developing novel ways of accessing the additional information that lies beyond the linear regime, using summary statistics that are sensitive to higher-order information, but yet impose minimal additional computational burden compared to a standard power spectrum evaluation. Among the long list of estimators of this kind that have been considered in the literature,[1] this active subfield involves proxy estimators [28–33], efforts to isolate the information encoded in the cosmic voids of the LSS [34–41], nonlinear transformations that partly restore the Gaussianity of the density field [17,42–47], splitting the density field into different environments [48–52], working with k-nearest neighbors [53,54] and a variety of other beyond 2-point statistics, such as Minkowski functionals [55–59], the minimum spanning tree [60] or 1-point statistics [61,62]. The recent rapid evolution of artificial intelligence (AI) has motivated efforts to extract cosmological non-Gaussianities using convolutional neural networks (CNNs) [63], demonstrating great promise in idealized settings [64–67]. Whether and how this simulation-based performance can be extended to reliable interpretations of actual galaxy data is still a matter of study; for example see Ref. [68].

Another path toward harnessing the nonlinear information encoded in the LSS, can be carved by seeking for a balanced trade-off between performance and interpretability, working in the middle-ground between traditional clustering estimators and CNNs. Such a trade-off is attempted by the wavelet scattering transform (WST), [63,69], which was first proposed in the context of computer vision. In direct analogy to the architecture of a CNN, a scattering network is constructed by successively performing two operations to an input field: wavelet convolution and modulus. After averaging over all pixels, the resulting outcome is a basis of interpretable WST coefficients, which can quantify the clustering information in the input field [70–72], while avoiding the previously discussed limitations of the standard moment expansion [17]. Motivated by these attractive properties, the WST has recently seen successful applications across the spectrum of natural sciences [73], including astrophysics [74–76], cosmology [77–85] and molecular chemistry [86,87].

As far as 3D clustering explorations are concerned, the first WST application was performed by Ref. [79], working with the fractional matter overdensity field obtained by N-body simulations [88] as input. Through a Fisher forecast, the basis of WST coefficients up to 2nd order was found to predict a substantial improvement on the 1-σ errors obtained on 6 cosmological parameters, exceeding the performance of both the standard and also the marked power spectrum. Another application was subsequently performed by Ref. [83], finding similar levels of improvement. Building upon these encouraging results, the subsequent work of Ref. [80] developed the first application of the WST to actual galaxy data, analyzing observations from the CMASS sample of the Baryon Oscillation Spectroscopic Survey (BOSS) [89,90] (under some approximations, however, as we explain in the next paragraph.) The WST, once again, was found to deliver a notable improvement to the errors obtained on 4 cosmological parameters, which were 3–6 times tighter compared to the ones from the galaxy power spectrum. This analysis demonstrated the great promise held in the use of the WST as a means of parameter inference in the context of spectroscopic surveys and precision cosmology in general.

Even though Ref. [80] laid out all the necessary steps to account for the complexities related to a WST application to spectroscopic galaxy data, combined with a set of high-fidelity galaxy mocks, it adopted a Taylor expansion approximation to model the cosmological dependence of the WST coefficients, which in principle could fail to capture non-Gaussianities present in the parameter likelihood. As a result, the accuracy of this approach was not tested in recovery tests against other simulations. In this work, we move beyond these approximations, and revisit our previous analysis with a full emulator predicting the cosmological dependence of the WST estimator. We take advantage of the full extent of the state-of-the-art suite of AbacusSummit simulations [91], which consists of a broad grid exploring variations in 8 cosmological parameters, in combination with a semianalytic model to parametrize the physics of galaxy formation for each cosmology. This extended suite enables the training of a neural net-based emulator that predicts the cosmological dependence of the WST coefficients in a 15-dimensional parameter space. In order to quantify the accuracy of this emulator, we subject it to a series of thorough parameter recovery tests against hold-out simulations, as well as against simulations using different models to capture small-scale galaxy physics. After we confirm that our model satisfies the necessary levels of accuracy for a reliable cosmological application, we use it to reanalyze the BOSS CMASS galaxy dataset, and obtain the marginalized 1-σ errors on 4 ΛCDM cosmological parameters, as well as on extended scenarios. We contrast our results against the ones obtained by the standard analysis performed using the multipoles of the anisotropic correlation function of galaxies, and discuss how our analysis compares to our previous work and prior ones in the literature.

[1]For a recent exploration of higher-order statistics in the particular context of weak lensing, also see Ref. [27].

Our paper is structured as follows: in Sec. II we introduce the wavelet scattering transform and in Sec. III we describe the BOSS dataset. We then proceed to lay out all the ingredients used to construct our simulation-based forward model in Sec. IV, as well as the details of our analysis pipeline in Sec. V. Finally, we present our results in Sec. VI, before concluding in Sec. VII. More technical details are discussed in Appendices A–D.

II. WAVELET SCATTERING TRANSFORM

The wavelet scattering transform [63,69] is a novel summary statistic that was proposed as an ideal middle-ground between a CNN and more traditional statistical estimators. Defined by a series of well-understood and interpretable mathematical operations, it can quantify the degree of clustering of an input field in a manner that not only matches, but also supersedes the properties of the standard 2-point correlation function [69].

According to the WST definition, the two fundamental properties to which an input field, $I(\mathbf{x})$, is subjected, are wavelet convolution and modulus. Specifically, given a localized wavelet probing a scale j_1 and an orientation l_1, that we will hereafter denote by $\psi_{j_1,l_1}(\mathbf{x})$, the WST transforms the input field as follows:

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

where $*$ indicates the convolution operation. If we further average over the transformed field in Eq. (1), we can derive a single number globally characterizing the field, which is called a WST coefficient. Furthermore, if the above sequence of elementary operations is successively repeated n times, and for a range of different j_1 scales and l_1 angles covered by a family of localized wavelets, $\psi_{j_1,l_1}(\mathbf{x})$, it will form a *scattering network*, with WST coefficients, S_n, given by:

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

explicitly shown up to order $n = 2$ above. The angular brackets, $\langle . \rangle$, in Eq. (2) and hereafter will denote taking the average value over the volume of the field.[2] Convolving with a localized wavelet essentially quantifies the strength of clustering in the input field over the relevant scales, similar to the 2-point function. The WST coefficients of order n have been shown to capture information related to the correlation function of order up to 2^n [69,71]. Building upon this property, it follows that the hierarchy of Eq. (2) leads to a collection of WST coefficients that can quantify the higher-order clustering information of the input physical field, $I(\mathbf{x})$, in analogy to the moment expansion usually applied to cosmological density fields. Opposite to the conventional series of correlation functions, however, the WST has been found to be more efficient at extracting information out of an input field, especially in highly non-Gaussian cases which are particularly challenging for higher-order moments to accurately describe [17,73]. Furthermore, the fact that the input field always enters Eq. (2) in a linear fashion guarantees a greater degree of numerical stability and robustness against outliers. In addition, the generated basis of WST coefficients is compact, such that the dimensionality of the resulting data vector can be kept under better control [73]. It is worth noting that the operations of wavelet (kernel) convolution, modulus (nonlinearity) and averaging (pooling), all implemented in a hierarchical scattering network, resemble the architecture and properties of a CNN with fixed kernels [63,69]. Combining all of the above properties, it becomes clear how the WST can be viewed as an interpretable alternative that lies between conventional summary statistics and CNNs, making it a potentially powerful tool to employ when harnessing higher-order information. In this work we will focus on the use of the WST for cosmological parameter inference, but we note that it can also be used in other applications, such as field synthesis and texture characterization, further discussed in Ref. [73].

Even though in the standard WST definition the input field enters Eq. (2) linearly, slightly relaxing this assumption and allowing for $I(\mathbf{x})$ to be raised to a power q, instead, results in the following variant:

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

which can lead to very interesting implications for cosmology, given that values of $q > 1$ or $q < 1$ respectively emphasize overdense or underdense regions of the LSS. This option was explored in the 3D matter overdensity WST application of Ref. [79], and was indeed found to produce more competitive constraints on cosmological parameters, with an emphasis on the sum of neutrino masses, when cosmic voids where highlighted using values of $q < 1$. In this application we will stay aligned with our previous work [80] and proceed with the version of WST given in Eq. (3).

Given that in this work we will focus on a WST application to 3D galaxy clustering, as we will specify below, the input field $I(\mathbf{x})$ will be taken to be 3-dimensional, even though the above discussion can in principle be valid for an arbitrary number of dimensions. Following our previous works [79,80], we adopt a mother wavelet given by the solid harmonic expression of

[2]Formally defined as the expectation value.

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

which was first applied in a 3D molecular chemistry application [86,87]. In Eq. (4), Y_l^m denote the usual Laplacian spherical harmonics and σ is the Gaussian width in units of the field grid size. The family of wavelets can then be generated by dilating the mother wavelet:

$$\psi_{j,l}^m(\mathbf{x}) = 2^{-3j}\psi_l^m(2^{-j}\mathbf{x}), \tag{5}$$

spanning different dyadic scales, 2^j, combined with varying values of the spherical harmonic of order l to describe the angular information of the wavelet family, after we sum over the remaining index m. In this case, the WST coefficients are then given by:

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

with

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

which is obtained using a 3D solid harmonic mother wavelet (5) in Eq. (3). We briefly note that other wavelets considered in the literature are Morlet wavelets [77,78], bump-steerable wavelets [81,83] or the equivariant wavelet construction of Ref. [75].

The total number of essential WST coefficients can be further reduced, compared to Eq. (3), if we notice that the second order scales $j_2 < j_1$, that is, scales smaller than the 1st order convolution scale j_1, are practically filtered out and do not carry any extra information. This fact was indeed confirmed in the 2D weak lensing (WL) application by Ref. [77] and was also subsequently adopted in our previous works [79,80]. We will also work with only one second order angular scale, that is, for $l_2 = l_1$ in Eq. (6), following the choice originally adopted by the solid harmonic implementation of Refs. [86,87]. Even though orientations $l_2 \neq l_1$ are expected to be informative, this choice has been shown to be a good trade-off [79,80,86,87] and will be adopted in this work as well.

The above choices determine the final number of produced WST coefficients, given as follows: for a certain total number of spatial scales J and harmonic angular orientations L, we will have:

$$(j, l) \in ([0, \ldots, J-1, J], [0, \ldots, L-1, L]), \tag{8}$$

giving rise to a total of

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

WST coefficients up to 2^{nd} order. Since the dilations of the mother wavelet scale are chosen to be dyadic, $J \leq \log_2(\text{NGRID})$, where NGRID is the resolution of the input field on each dimension. Finally, the width, σ, of the Gaussian in Eq. (4) and the power, q, in Eq. (6) are free parameters, whose values will be determined in the next section for our particular galaxy clustering application.

To summarize, for a given choice of J, L, q and σ, an input field $I(\mathbf{x})$ of resolution NGRID^3 gives rise to the WST coefficients (9) evaluated from Eq. (6). We perform this evaluation using the publicly available package KYMATIO [92],[3] as we will explain in the next section.

III. DATASET

In this section we introduce the dataset that will be analyzed in this work. This consists of luminous red galaxies (LRGs) obtained from the twelfth data release (DR12) [93] of the Baryon Oscillation Spectroscopic Survey (BOSS), a part of Sloan Digital Sky Survey, SDSS-III [89,90], in particular the CMASS sample.[4] Following our previous application [80], which was in turn aligned with the original analyses of BOSS data [94,95], we will work with each of the two subsamples obtained in the Northern (NGC) and the Southern Galactic Cap (SGC). If $X_{\rm NGC}$ and $X_{\rm SGC}$ denote the summary statistics evaluated from the Northern and Southern parts of the BOSS footprint, with angular area equal to $A_{\rm NGC} = 6851$ deg^2 and $A_{\rm SGC} = 2525$ deg^2, respectively, then we will always work with the weighted average

$$X_{\rm N+S} = \frac{(A_{\rm NGC}X_{\rm NGC} + A_{\rm SGC}X_{\rm SGC})}{(A_{\rm NGC} + A_{\rm SGC})}, \tag{10}$$

where X in our analysis will be the data vector of the WST coefficients or the multipoles of the anisotropic correlation function of galaxies, as we will further explain in Secs. IV and V. Furthermore, we identify and work with the part of the sample with galaxy number density greater than $3 \times 10^{-4}\ h^3/\text{Mpc}^3$, corresponding to the redshift range $0.4613 < z < 0.5692$. In order to generate a sample with a

[3]Available in https://www.kymat.io/. We clarify that 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 exactly equivalent for parameter inference applications. We follow this version and, strictly speaking, we work with the sum over all pixels rather than the mean.

[4]All data are publicly available at https://data.sdss.org/sas/dr12/boss/lss/.

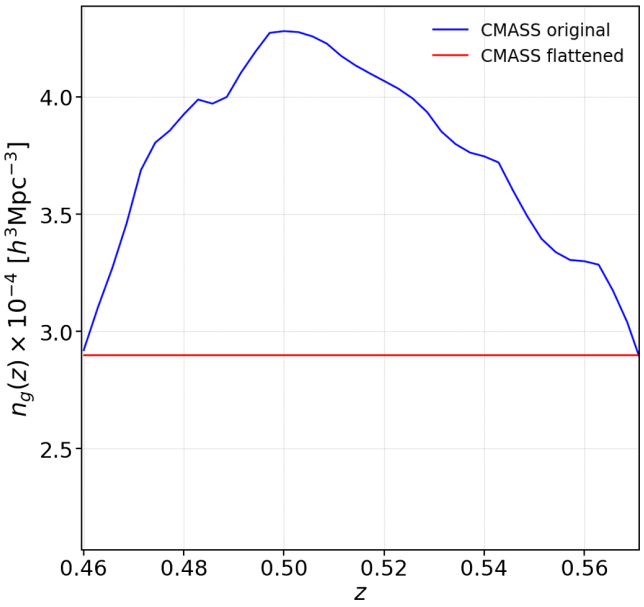

FIG. 1. The galaxy number density of the original CMASS sample as a function of redshift z (blue), shown with the final downsampled version of constant number density, $\bar{n}_g = 2.9 \times 10^{-4}\ h^3/\mathrm{Mpc}^3$ (red), that we work with in order to match the constant density profile of the AbacusSummit mocks.

constant density profile as a function of redshfit z, we further bin these galaxies into 50 linearly spaced z bins, and randomly downsample each bin such that the final outcome is a sample with a constant galaxy number density $\bar{n}_g = 2.9 \times 10^{-4}\ h^3/\mathrm{Mpc}^3$. In Fig. 1, we show the original varying $n_g(z)$ of the sample, together with the final flattened profile that we are going to work with. The choice of the target density $\bar{n}_g = 2.9 \times 10^{-4}\ h^3/\mathrm{Mpc}^3$ is motivated by the mocks we use, and will be further explained in Sec. IV. We also note that the choice to work with a flat density profile, which is meant to ensure a more accurate modeling of our density-dependent WST estimator, is different than our previous analysis [80], in which we worked with the original varying $n_g(z)$ in the range $0.46 < z < 0.60$. A similar choice has been made in other recent simulation-based reanalyses of BOSS data [96,97].

IV. SIMULATION-BASED FORWARD MODEL

In this section we will describe the various ingredients used to construct our simulation-based model for the galaxy clustering and its summary statistics as a function of the cosmological and galaxy model parameters of interest. We begin with the suite of the AbacusSummit simulations used for the nonlinear modeling of the dark matter density and velocity fields, and then introduce the semi-analytical AbacusHOD framework for populating the gravitationally bound dark matter halos with galaxies. Finally, we explain how we evaluate the WST coefficients and the 2-point correlation function multipoles of the galaxy mocks, in order to construct the training set for our emulator.

A. The AbacusSummit simulations

AbacusSummit [91] is a suite of state-of-the-art cosmological N-body simulations that were run with the Abacus N-body code [98,99]. Containing more than 150 high-accuracy and high-resolution simulations spanning almost 100 different cosmologies, it is capable of not only matching but also exceeding the simulation requirements of the Dark Energy Spectroscopic Instrument (DESI) survey [2,100]. As a result, it is the ideal set to use in order to produce high-fidelity galaxy mocks for our BOSS CMASS simulation-based reanalysis. We will exclusively work with the main ("base") set of cubic boxes with a side of length 2 Gpc/h, that evolved 6912^3 dark matter particles with an individual mass equal to $2.1 \times 10^9\ h^{-1}M_\odot$.

In order to identify gravitationally bound dark matter halos in the simulations, the AbacusSummit uses a new efficient spherical-overdensity (SO) halo finder called CompaSO [101], which performs this task on-the-fly, and includes a series of improvements to avoid previously known challenges faced by halo finders, such as failure to identify structures close to larger halo centers or the blending of halos. Further details about Abacus and CompaSO can be found in the corresponding papers referenced above.

1. The cosmology grid

Our cosmology grid consists of 85 simulations performed for different variations in the values of 8 cosmological parameters, which form the basis of our emulator and parameter inference setup. These parameters are: the baryon density $\omega_b = \Omega_b h^2$, the cold dark matter density $\omega_{\mathrm{cdm}} = \Omega_{\mathrm{cdm}} h^2$, the rms amplitude of linear density fluctuations at 8 Mpc/h σ_8, the spectral tilt n_s, the running of the spectral tilt α_{run}, the effective number of relativistic degrees of freedom N_{eff}, and the dark energy equation of state parameters w_0 and w_a ($w(a) = w_0 + (1-a)w_a$), where $h = H_0/(100\ \mathrm{km\,s^{-1}\,Mpc^{-1}})$ is the dimensionless Hubble constant. Each one of the 85 simulations has been performed for the same fixed initial random phase, and with the value of the Hubble constant, H_0, chosen such that the comoving angular size of the sound horizon at last scattering, $\theta_\star$, is fixed to the corresponding value derived from measurements by the *Planck* satellite [102], $100\theta_\star = 1.041533$.

We refer to the different cosmologies using the naming scheme cXXX, where XXX ranges from 000 to 181. Details for each one of them are presented in the AbacusSummit website,[5] with a visualization of the cosmological

[5]https://abacussummit.readthedocs.io/en/latest/cosmologies.html.

TABLE I. Priors bounds used to generate the cosmology + HOD training set of our emulator. Units of mass are in $h^{-1}M_\odot$. The HOD values are roughly centered on results from Ref. [103].

Parameter	Bounds
ω_b	[0.0207, 0.0243]
ω_c	[0.1032, 0.14]
σ_8	[0.687, 0.938]
n_s	[0.901, 1.025]
$a_{\rm run}$	$[-0.038, 0.038]$
$N_{\rm eff}$	[2.1902, 3.9022]
w_0	$[-1.27, -0.70]$
w_a	$[-0.628, 0.621]$
$\log_{10} M_{\rm cut}$	[12.4, 13.3]
$\log_{10} M_1$	[13.0, 15.0]
σ	[0.001, 1.0]
α	[0.5, 1.5]
κ	[0.0, 8]
$\alpha_{\rm c}$	[0.0, 0.8]
$\alpha_{\rm s}$	[0.0, 1.5]

parameter grid shown in Fig. 1 of Ref. [103] and its bounds listed in Table I.

We briefly describe the specifications of some selected cosmologies contained in the parameter grid. `c000` is a ΛCDM cosmology that corresponds to the parameters inferred by the *Planck* 2018 [102] TT, TE, EE + lowE + lensing likelihood analysis, and which we pick as our fiducial cosmology from now on.

Furthermore, there are four secondary cosmologies exploring variations around the fiducial, `c001-004`, which we will use to validate the accuracy of our emulator in the next section. `c001` corresponds to the WMAP9 + ACT + SPT cosmology [104], while `c002` is a wCDM cosmology with $w_0 = -0.7$ and $w_a = -0.5$. Finally, `C003` is a cosmology with higher $N_{\rm eff} = 3.7$, and `c004` has a low clustering amplitude given by $\sigma_8 = 0.75$.

The AbacusSummit also contains additional simulations that vary each one of the 8 cosmological parameters, in turn, and in a step-wise fashion around the fiducial `c000`, while keeping the rest fixed, in order to enable the evaluation of first-order derivatives for summary statistics. This linear derivative grid consists of cosmologies `c100-126`, which were used to construct the Taylor expansion approximation we adopted in Ref. [80]. Cosmologies `c130-181` form a broad parameter grid that provides a wider coverage of the 8-dimensional target parameter space and enables the training of emulators. Further details on the motivations behind the choice of these cosmologies and the parameter ranges can be found in Ref. [91] and the AbacusSummit website.

Lastly, in order to quantify the effects of sample variance and potential errors introduced when training at a single phase, a second set of simulations with the same specifications has been run for 24 additional random realizations of the `c000` fiducial cosmology. The phase information is labeled as ph000-024. In the next sections we will describe how we used both of the above sets in order to accurately train our emulator for the vector of WST coefficients and the multipoles of the 2-point correlation function.

B. The halo occupation distribution (HOD)

The galaxy–halo connection model we use to generate the galaxy mocks for our forward model is known as the halo occupation distribution (HOD) (see, e.g., Refs. [105,106]), which is a probabilistic model that populates dark matter halos with galaxies through a set of empirical formulas conditioned on halo properties. For a luminous red galaxy (LRG) sample such as CMASS, the HOD is well approximated by a vanilla model given by:

$$\bar{n}^{\rm LRG}_{\rm cent}(M) = \frac{1}{2}\,{\rm erfc}\left[\frac{\log_{10}(M_{\rm cut}/M)}{\sqrt{2}\sigma}\right], \tag{11}$$

$$\bar{n}^{\rm LRG}_{\rm sat}(M) = \left[\frac{M - \kappa M_{\rm cut}}{M_1}\right]^{\alpha} \bar{n}^{\rm LRG}_{\rm cent}(M), \tag{12}$$

where the five parameters characterizing the model are $M_{\rm cut}, M_1, \sigma, \alpha, \kappa$. The parameter $M_{\rm cut}$ defines the minimum halo mass to host a central galaxy, M_1 sets the typical halo mass that hosts one satellite galaxy, σ characterizes the steepness of the error function upturn in the number of central galaxies, α is the power-law index on the number of satellite galaxies, and $\kappa M_{\rm cut}$ controls the minimum mass of a halo that can host a satellite galaxy. We have also added a modulation term $\bar{n}^{\rm LRG}_{\rm cent}(M)$ to the satellite occupation function to disfavor satellites from halos without centrals. This term represents a model choice and is inconsequential for the conclusions of this work.

The HOD model does not only provide predictions for the number of galaxies populating each halo, but it also determines the positions and velocities of these galaxies. In the case of the central galaxies, their positions and velocities match the ones of the halo center-of-mass (the L2 subhalo when working with CompaSO), while the satellites are randomly assigned to halo particles with uniform weights, each satellite inheriting the position and velocity of its host particle. Note that we do not impose any satellite radial profile in this model.

We also include a motivated HOD extension known as velocity bias, which biases the velocities of the central and satellite galaxies relative to their host halos and particles. This is shown to be a necessary ingredient in modeling BOSS LRG redshift-space clustering on small scales [e.g. [107,108]]. The velocity bias has also been identified in hydrodynamical simulations and measured to be consistent with observational constraints [e.g. [109,110]].

We parametrize the velocity bias through two additional HOD parameters:

(i) $\alpha_{\rm vel,c}$ controls the peculiar velocity of a central galaxy relative to the halo center, and is called the central velocity bias parameter. For instance, a value of $\alpha_{\rm vel,c} = 0$ indicates that centrals perfectly track the velocity of halo centers.

(ii) $\alpha_{\rm vel,s}$, the satellite velocity bias, is the equivalent parameter for the satellite galaxies, modulating how their peculiar velocities deviate from those of the local dark matter particles. A value of $\alpha_{\rm vel,s} = 1$ indicates that satellites perfectly track the velocity of the underlying dark matter particles.

Furthermore, we do not include the effects of assembly bias in our analysis, given that they typically manifest in smaller scales than the ones we consider, as we clarify below. We additionally check our cosmology recovery against a galaxy mock that contains galaxy assembly bias in Sec. V B. Nevertheless, we acknowledge the lack of robust galaxy assembly bias modeling as a potential systematic. We reserve an analysis extending to smaller scales for a follow-up investigation. For a detailed discussion on the effects on assembly bias on cosmological analyses of BOSS CMASS, readers are referred to Ref. [111].

For computational efficiency, we adopt the highly optimized AbacusHOD implementation, which significantly speeds up the HOD calculation per HOD parameter combination [108]. The code is publicly available as a part of the ABACUSUTILS package at [112]. Example usage can be found at [113]. In order to match the clustering of CMASS in the redshift range $0.4613 < z < 0.5692$, we produce cubic galaxy mocks (of side 2 Gpc/h) at redshift $z = 0.5$.

To summarize, the HOD model used in this analysis is fully parameterized by 7 parameters, $M_{\rm cut}$, M_1, α, $\alpha_{\rm vel,c}$, $\alpha_{\rm vel,s}$, κ and σ.

C. Survey geometry

The AbacusSummit galaxy mocks that we produce with AbacusHOD come in a periodic cubic geometry with a side equal to 2 h^{-1}Gpc, at output redshift $z = 0.5$, as we previously discussed. This configuration is different from the nontrivial survey geometry of the CMASS sample that we will analyze in this work, which was introduced in Sec. III. When working with conventional statistics, the effect of a nontrivial survey geometry can be usually captured with a model. In the case of the galaxy power spectrum, for example, the prediction for a periodic configuration is convolved with the Fourier transform of the survey mask [94,95,114–116] or, equivalently, the prediction from the masked data can be de-convolved [117]. Given that no such model is available for the WST, which is sensitive to the survey geometry through the successive wavelet convolutions, we proceed to directly cut the Abacus cubes into the exact 3D shape of the BOSS CMASS data, as we did in our previous work [80].[6] Specifically, each cubic mock of redshift $z = 0.5$ is downsampled to a constant number density $\bar{n}_g = 2.9 \times 10^{-4}\ h^3/{\rm Mpc}^3$ and is then fed as input into the public code MAKE_SURVEY [119].[7] Using the real-space Cartesian positions and velocities for each galaxy at $z = 0.5$, the CMASS angular footprint, as well as the parameters for each cosmology cXXX, MAKE_SURVEY transforms the original cubic mocks into galaxy catalogs with sky coordinates right ascension (RA), declination (DEC), and redshift z that exactly match the 3D geometry of the observed CMASS sample in the target range $0.4613 < z < 0.5692$, with the redshift-space distortion (RSD) implemented along the radial direction. The procedure is repeated twice for each mock in order to produce separate samples for NGC and SGC, respectively. As in Ref. [80], we confirm the robustness of this procedure by evaluating the power spectra of both the original cubic and the final cut-sky mocks and by making sure they remain unchanged, up to sample variance error.

D. Summary statistic evaluation

Having laid out the procedure to generate realistic galaxy mocks that resemble the footprint of the CMASS sample as a function of the cosmological and HOD parameters, we now proceed to explain how we evaluate the summary statistics of interest, starting with the WST coefficients.

1. WST

The quantity of interest for the density-dependent WST estimator is the fractional overdensity field of galaxies, which we evaluate with the following procedure: the sky coordinates RA, DEC and z of each galaxy in each sample (be it either the cut-sky mocks or the CMASS data) are converted to comoving Cartesian coordinates (x,y,z), always assuming a fiducial flat ΛCDM cosmology with $\Omega_m = 0.3152$, $h = 0.6736$ (corresponding to the Abacus cosmology c000). Each sample is then embedded into the smallest possible 3D cube for this task, which we determine with the public package NBODYKIT,[8] and which is found to have a comoving side $L = 2700$ Mpc/h for the range $0.4613 < z < 0.5692$. When working with spectroscopic data in sky coordinates, the relevant quantity is the (weighted) fractional overdensity of galaxies, also known as the Feldman-Kaiser-Peacock (FKP) field, $F(r)$ [114], given by:

$$F(\mathbf{r}) = \frac{w_{\rm FKP}(\mathbf{r})}{I_2^{1/2}}[w_c(\mathbf{r})n_g(\mathbf{r}) - \alpha_r n_s(\mathbf{r})], \qquad (13)$$

[6]Alternatively, one could consider using modern inpainting techniques [118].

[7]Available at https://github.com/mockFactory/make_survey.

[8]https://nbodykit.readthedocs.io/en/latest/index.html.

which can be evaluated on a 3D Cartesian grid. The quantities $n_g(\mathbf{r})$ and $n_s(\mathbf{r})$ in Eq. (13) denote the observed number density of the galaxies compared to the one of a random, unclustered, catalog, respectively, with the latter containing α_r times more objects. Furthermore, the BOSS dataset is accompanied by a set of systematic weights given by Refs. [94,95]:

$$w_c(\mathbf{r}) = (w_{\rm rf}(\mathbf{r}) + w_{\rm fc}(\mathbf{r}) - 1.0) w_{\rm sys}(\mathbf{r}), \tag{14}$$

in which a fiber collision weight, $w_{\rm fc}$, a systematics weight, $w_{\rm sys}$ and a redshift failure weight, $w_{\rm rf}$, are combined to account for the various incompletenesses of the observed sample. Aiming to ensure optimal recovery of small-scale information from the galaxy power spectrum, we traditionally also define the FKP weight [114]:

$$w_{\rm FKP}(\mathbf{r}) = [1 + \bar{n}_g(\mathbf{r}) P_0]^{-1}, \tag{15}$$

where $P_0 = 10^{-4}\ {\rm Mpc}^3/h^3$, and where the normalization factor

$$I_2 = \int d^3\mathbf{r}\, w_{\rm FKP}^2(\mathbf{r}) \langle w_c(\mathbf{r}) n_g(\mathbf{r})\rangle^2 \tag{16}$$

is defined in Eq. (13), with respect to the amplitude of the regular galaxy power spectrum of a uniform sample. In addition to the values of the systematic weights (14) for each observed galaxy, the public BOSS release also includes random catalogs matching the same selection function and footprint of the survey, in order to enable the evaluation of $n_s(\mathbf{r})$ in Eq. (13). Out of the various options available, we choose to work with the random catalog corresponding to $\alpha_r = 50$, which is the commonly adopted choice in the literature [94,95]. When working with a sample that does not possess incompleteness weights, as is the case for the galaxy mocks, Eq. (13) merely corresponds to the regular galaxy overdensity field, but in a nontrivial survey geometry. In order to evaluate the random density field $n_s(\mathbf{r})$ in this case, we similarly generate a random cubic sample with 50 times higher number density than the original mocks, and then subject it to the same cut-sky procedure that we described in the previous subsection.

We should note, at this point, that in order to convert the sky coordinates of the galaxies in the CMASS sample into comoving Cartesian ones, we assumed a (potentially incorrect) flat ΛCDM cosmology corresponding to $\Omega_m = 0.3152$, $h = 0.6736$. This assumption introduces an error usually referred to as the Alcock–Paczynski (AP) distortion [120]. To account for this effect in our model, we always assume the above same cosmology when converting the coordinates RA, DEC and z of the Abacus mocks back into comoving ones, even though the true cosmological parameters for each one of the cXXX boxes is actually known (and were used to convert the original cubes into cut sky mocks). The procedure is the equivalent one to the power spectrum rescalings usually applied in order to account for the AP effect in traditional BOSS analyses [23,97,121–125], that we also adopted in Ref. [80].

Finally, Eq. (13) with the corresponding systematic weights from Eq. (14) (or unweighted) and the FKP weight (15), can be combined with the random catalogs in order to enable the evaluation of the final FKP density field from the CMASS data (Abacus mocks). Following the choices adopted in our previous WST BOSS analysis [80], we resolve the field on a mesh of resolution NGRID = 270, using the triangular shaped cloud (TSC) mass assignment scheme [126], and work with a Gaussian width $\sigma = 0.8$ in Eq. (6), such that the smallest density cell corresponds to a scale of length 8 h^{-1} Mpc on the side. This FKP field serves as input into the WST network (6) in order to evaluate the relevant WST coefficients. Combining the above choices with $J = 4$ scales, $L = 4$ orientations, and $q = 0.8$ (as in Ref. [80]), we obtain the target data vector of 76 WST coefficients from Eq. (6). The evaluation is performed with KYMATIO [92], using our modified version for an application to a masked galaxy field (as explained further in Appendix A of Ref. [80]). We note that the overall evaluation of the WST coefficients out of an original Abacus cubic mock through the pipeline described above takes about 60 seconds per core when the WST evaluation is GPU-accelerated.

2. 2-point correlation function

In order to have a benchmark that will allow us to assess the performance of the WST compared to standard cosmological analyses, we also evaluate the 2-point correlation function of galaxies. In particular, if by $\xi(s, \mu_s)$ we denote the 2D anisotropic correlation function of galaxies as a function of redshift space separation s, then its multipoles, $\xi_\ell(s)$, can be extracted through the usual expansion

$$\xi(s, \mu_s) = \sum_l \xi_\ell(s) L_\ell(\mu_s) \tag{17}$$

in a basis of Legendre Polynomials $L_\ell(\mu_s)$, which then gives

$$\xi_\ell(s) = (2\ell + 1) \int_0^1 \xi(s, \mu_s) L_\ell(\mu_s) d\mu_s. \tag{18}$$

For a sample of galaxies in sky coordinates, which we are working with in this analysis, $\mu_s = \hat{s} \cdot \hat{r}$, where the radial anisotropy direction $\hat{r}$ is the line-of-sight (as opposed to one of the Cartesian axes direction when working with a periodic box). We choose to work with the two lowest nonvanishing multipoles, $\ell = \{0, 2\}$, which we evaluate with the public code Pycorr,[9] which is a wrapper for Corrfunc [127],[10] using the Landy-Szalay (LS) estimator [128] with 241 linearly spaced angular bins in $-1 < \mu < 1$. For the

[9] https://github.com/cosmodesi/pycorr.
[10] https://corrfunc.readthedocs.io/en/master/index.html.

spatial separation, we adopt a differential binning strategy, which is the following: for the monopole, we use 10 linearly spaced bins centered between $10 < s < 56$ Mpc/h followed by 6 bins for scales $67 < s < 142$ Mpc/h, while for the quadrupole we downsample the above binning scheme by a factor of 2. This choice, which corresponds to a total of 24 bins, was found to deliver the optimal trade-off between large-scale noise due to cosmic variance and the ability to capture the full shape of the correlation function. We also note that this binning scheme is still finer that the one chosen for the WST, in order to ensure a fair comparison between the performance of the two statistics.

This evaluation can be straightforwardly performed using the sky positions of the CMASS galaxy sample (or the simulated mocks), as well as the ones of the accompanied random catalogs, as input to Corrfunc. For the conversion of the sky coordinates into comoving ones, we adopted the same fiducial cosmology as discussed in Sec. IV D 1 for the WST, in connection to the AP effect. The choice of $s_{\rm min}$ matches the minimum scale accessed by the WST, for which we used a cell of grid size 8 Mpc/h as explained above, in order to ensure a fair comparison. (Further discussion on the minimum scale cut can be found in Appendix E).

E. Emulator

After explaining the steps to go from the original AbacusSummit simulations to realistic galaxy mocks resembling the properties of the CMASS sample, we now lay out the details of our emulation scheme for the cosmological dependence of the target summary statistics.

Emulators refer to parametrized surrogate models for the cosmological dependence of a summary statistic used to interpolate sparse likelihood evaluations. The emulator replaces the expensive likelihood calls with the much cheaper emulator model calls, thus enabling a much faster sampling at the cost of introducing additional errors in the model training. Such schemes have become increasingly popular in simulation-based cosmological analyses with the advent of fast yet flexible machine learning models such as neural nets and Gaussian processes, with a series of successful cosmology applications in recent years (see e.g. Refs. [97,103,129–136]).

To generate the training and test set, we forward model the final summary statistics (WST coefficients and 2-point correlation function) across 85 cosmologies and 2700 HOD variations at each cosmology, creating an initial set of 229500 mocks. The cosmology grid is described in Sec. IV A 1 and spans the wCDM + $N_{\rm eff}$ + running space around *Planck* 2018 values [102]. We leave out the four secondary cosmologies C001-004 as out-sample tests. The HODs are sampled in a Latin Hypercube with flat bounded priors along each HOD parameter direction. The bounds for all parameters are summarized in Table I. For each cosmology and HOD, we generate the periodic galaxy mocks according to the steps described in Secs. IV A and IV B, discard the mocks that have number density lower than $2.9 \times 10^{-4}\ h^3/\mathrm{Mpc}^3$, and randomly downsample the galaxies of the other mocks in order to exactly match the target density $\bar{n}_g = 2.9 \times 10^{-4}\ h^3/\mathrm{Mpc}^3$. The value of this density cut-off allows us to retain a significant portion of the original collection of 229500 mocks, while discarding HOD configurations resulting in very low number densities, as shown in Fig. 2. We end up retaining 151474 cubic mocks with number density $\bar{n}_g = 2.9 \times 10^{-4}\ h^3/\mathrm{Mpc}^3$, each one of which is cut to give two independent cut sky galaxy samples for NGC and SGC, respectively, as explained in Sec. IV C. We extract the summary statistics (WST and 2-point function) out of each one of them, as explained in Sec. IV D, and finally obtain the corresponding sky-averaged quantities according to Eq. (10), which form our final emulator training + test set.

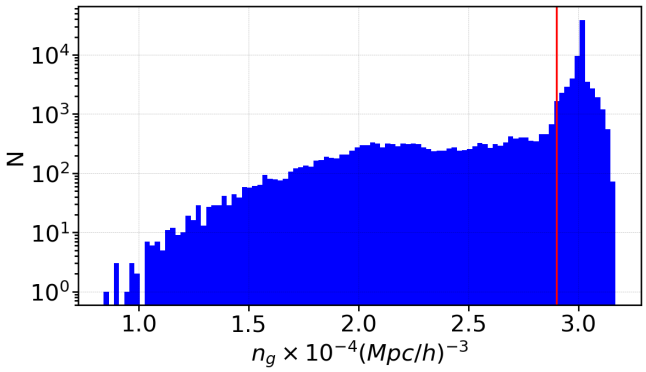

FIG. 2. A histogram of the distribution of galaxy number densities of the galaxy mocks forming our original emulator parameter grid. Mocks with number densities lower than the cut-off $\bar{n}_g = 2.9 \times 10^{-4}\ h^3/\mathrm{Mpc}^3$ (red vertical line) are discarded, while the rest are downsampled to exactly match this value, in order to ensure a robust modeling of the density-dependent WST estimator since this is the constant density value used in the Abacussummit mocks. The outcome of this procedure forms the final emulator training set consisting of 151474 mocks, i.e. those lying on the right of the red vertical line in the histogram.

For the emulator, we adopt a fully connected neural network as our surrogate model. For the emulation of WST, we adopt a network of 3 layers as our fiducial model, with 300 nodes in each layer and a Gaussian error linear unit (GELU) activation function. We train the network with the Adam optimizer and a mean squared loss function taking the diagonal terms of the CMASS WST covariance matrix (the evaluation of which is explained in Sec. V A) as weights. We follow a minibatch procedure and conduct cross-validation throughout the training process.

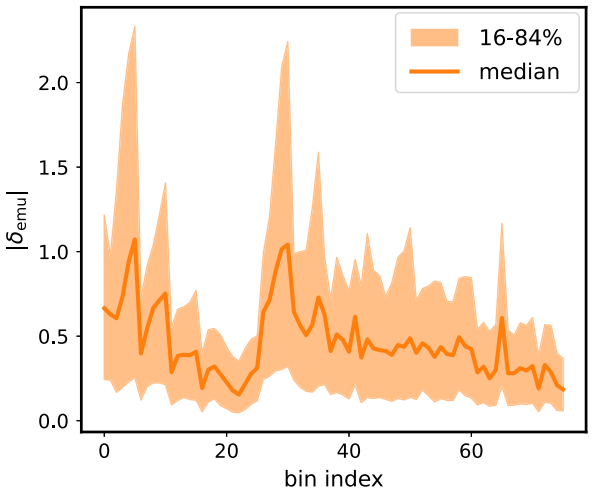

FIG. 3. The median WST emulator error tested on the four leave-out cosmologies as a function of bin index of the WST coefficients vector. The y-axis denotes the emulator error normalized by the CMASS 1σ uncertainty.

We visualize the final WST emulator performance in Figs. 3 and 4. Specifically, Fig. 3 summarizes the emulator error relative to the CMASS uncertainty, $\delta_{\rm emu}$, as a function of WST bin indices. The errors are computed on 1000 HODs (sampled from the prior) at the four out-sample test cosmologies. The bins with the largest relative errors are the ones probing the largest spatial scales, which are more susceptible to cosmic variance and thus exhibit a larger dispersion in their values. We report a mean $|\delta_{\rm emu}|$ of 0.51, suggesting that the emulator error is overall sub-dominant relative to the measurement uncertainties. Figure 4 compares the true WST values and the emulator predicted values across all test cases for a few selected WST coefficients of the data vector (i.e. bins). The orange points show the emulator predictions for the respective coefficients for each one of the 4000 leave-out test cases, whereas the blue band shows the measurement uncertainties. The dashed line shows the $Y_{\rm pred} = Y_{\rm true}$ line. We see no sign of bias in the emulator prediction. Lastly, we repeat the above steps for the WST in order to create the corresponding emulator for the multipoles of the 2-point correlation function using the same training set.

V. ANALYSIS

In this section, we lay out the details of how we will use our forward model for the galaxy clustering in order to perform a likelihood analysis of the BOSS data. We start with a description of the adopted likelihood we sample

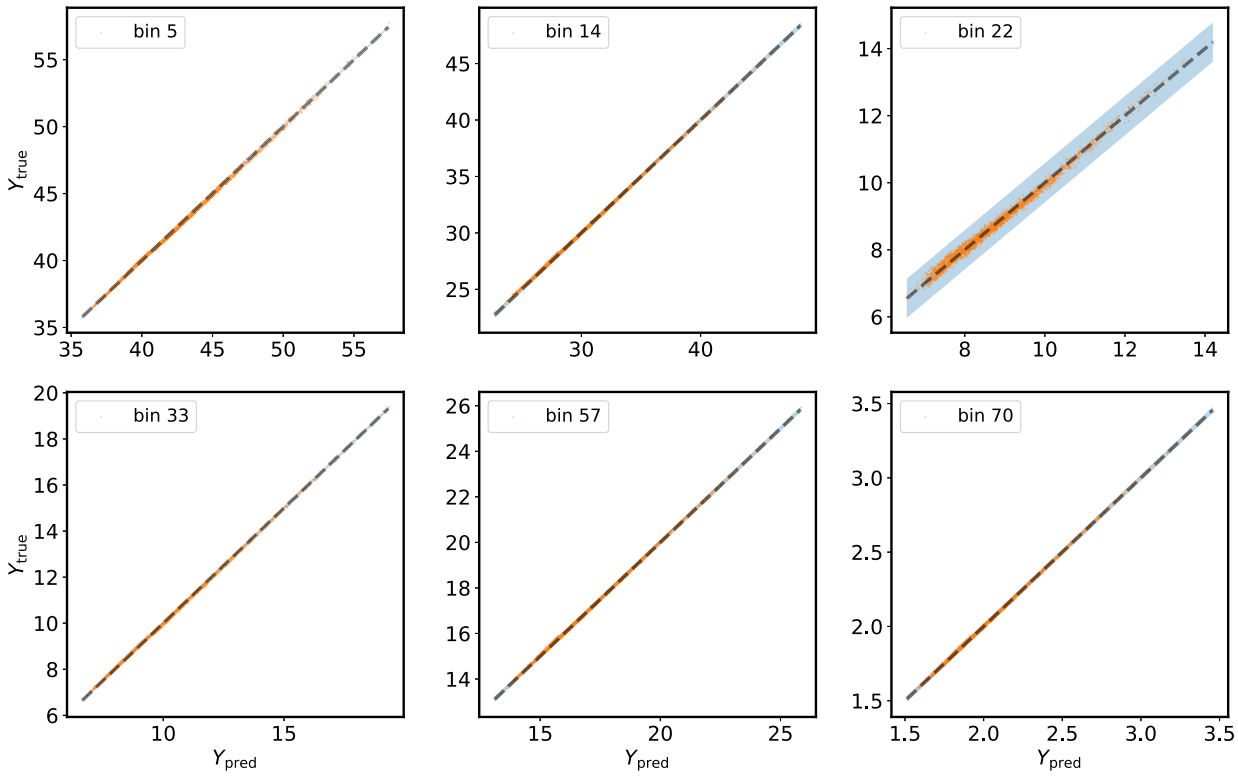

FIG. 4. The bias of the WST emulator tested on the four leave-out cosmologies for six randomly selected coefficients (bins) of the WST data vector. The legend shows the WST bin indices. The orange scatter points showcase the true and predicted values of the WST coefficients for each one of the 4000 leave-out tests, whereas the blue band corresponds to the 1σ uncertainty of the CMASS WST measurement.

from and then explain our steps to validate the robustness of our pipeline.

A. Likelihood modeling

Having laid out our methodology on how to forward model the cosmological dependence of the WST coefficients, as well as on how to extract the corresponding prediction from the data, we now explain our strategy for combining these necessary ingredients to perform a likelihood analysis of the BOSS dataset. Consider $\mathbf{X}$ to be the summary statistic of interest, which in our analysis denotes either the vector of WST coefficients or the multipoles of the correlation function (or their combination). Assuming $\mathbf{X}$ is Gaussian-distributed, as we confirm in Appendix A, its likelihood, $\mathcal{L}(\theta|\mathbf{d})$, will then follow the familiar form:

$$\log\mathcal{L}(\theta|\mathbf{d}) = -\frac{1}{2}[\mathbf{X_d} - \mathbf{X}_t(\theta)]^{\mathrm{T}} C^{-1} [\mathbf{X_d} - \mathbf{X}_t(\theta)] + \text{const}, \tag{19}$$

with $\mathbf{X_d}$ being the value of the estimator evaluated from the BOSS CMASS dataset $\mathbf{d}$, that we will analyze in order to infer the set of parameters θ. Furthermore, C in Eq. (19) denotes the covariance matrix of $\mathbf{X}$, which can be decomposed as in Ref. [103]:

$$C = C_{\mathbf{d}} + C_{\mathrm{emu}} + C_{\mathrm{phase}}. \tag{20}$$

The first term in Eq. (20), $C_{\mathbf{d}}$, represents the usual contribution from the sample variance of the CMASS dataset $\mathbf{d}$, given by:

$$C_{\mathbf{d}} = \frac{1}{\mathrm{N_{mocks}} - 1} \sum_{\mathrm{k}=1}^{\mathrm{N_{mocks}}} (\mathbf{X}_P^k - \bar{\mathbf{X}}_P)(\mathbf{X}_P^k - \bar{\mathbf{X}}_P)^{\mathrm{T}}, \tag{21}$$

which we evaluate using $N_{\mathrm{mocks}} = 2048$ realizations of the Patchy mocks (to be described in Sec. V A 1), and with $\bar{\mathbf{X}}_P$ being the mean prediction over the N_{mocks}. Furthermore, we follow Ref. [103] and consider two extra contributions to the overall error budget, which reflect additional sources of uncertainty arising from our forward model and are essential for a reliable interpretation of our analysis. In particular, C_{emu} quantifies the residual emulator error evaluated (at fixed phase ph000) by averaging over the $4 \times 1000 = 4000$ hold-out test errors, generated from `c001-c004` (as introduced in Sec. IV A). These hold-out tests and their results will be described in detail in Sec. V B.

Furthermore, C_{phase} is meant to capture the effect of training using mocks at a fixed phase, rather than the average over many random realizations. To mitigate this effect, we make use of 24 additional simulations initialized at phases ph001-ph024, for the fiducial `c000` cosmology and a fixed HOD (corresponding to the best-fit values from [103]). We then apply the following phase correction to the data vector:

$$\mathbf{X}_{\mathrm{smooth}} = \mathbf{X}_{\mathrm{ph000}} \left[\frac{\bar{\mathbf{X}}}{\mathbf{X}_{\mathrm{ph000}}}\right], \tag{22}$$

where $\mathbf{X}_{\mathrm{ph000}}$ is the original emulator prediction, trained at fixed phase, and the term inside the brackets denotes the fractional correction evaluated over the 25 random realizations for the cosmology `c000`. Even though Eq. (22) assumes that this phase effect is cosmology-independent, it was found in Ref. [103] to be sufficient for the mitigation of cosmic variance to the emulator predictions, and we adopt it here as well. Equivalently, one could explicitly evaluate an error term, C_{phase}, using the 25 random phase realizations, as in Eq. (21). We have tried both approaches and have found minimal differences between the results of the corresponding likelihood analyses. It is straightforward to see that in the limit of perfect emulator accuracy, these two additional terms would vanish, and Eq. (20) would reduce to its usual expression capturing only the cosmic variance (21), but we will find that these effects are not negligible.

Furthermore, upon inversion of the covariance matrix in Eq. (19), we apply the standard debiasing Hartlap correction factor [137]:

$$\hat{C}^{-1} = \frac{N_{\mathrm{mocks}} - N_{\mathbf{d}} - 2}{N_{\mathrm{mocks}} - 1} C^{-1}, \tag{23}$$

where $N_{\mathbf{d}}$ is the dimensionality of $\mathbf{X_d}$, which will be $N_{\mathbf{d}} = 76$ for the WST coefficients, $N_{\mathbf{d}} = 24$ when working with the $l = 0, 2$ multipoles of the correlation function (down to $r_{\mathrm{min}} = 10.5$ Mpc/h), and $N_{\mathbf{d}} = 100$ for the joint analysis. Before inverting, 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. (19), and also that the number of realizations is sufficient for them to be well-converged (a very similar test for this can be found in Appendix B of Ref. [80]). The correlation matrix, $C_{ij}/(C_{ii}C_{jj})$, of the joint statistic consisting of the 2-point function multipoles and the WST coefficients is shown in Fig. 5, evaluated at the fiducial cosmology. Focusing on the WST coefficients on the upper right subplot, and starting with the 1st order group of wavelets (that is, until index 25), we notice the existence of strong correlations between nearby scales and angles (close to the diagonal), which progressively decrease and turn into anticorrelations between the smallest and the largest wavelet scales. Similar patterns permeate into the 2nd order group of wavelets and their correlations with the corresponding 1st order scales. The correlation matrix of the 2-point correlation function multipoles, corresponding to the lower left corner, exhibits the familiar structure known in the literature [138]. Lastly, when looking into the joint covariance between the two statistics in Fig. 5, we observe the existence of positive correlations between the wavelets and the 2-point function monopole, which are most pronounced with the wavelets probing the largest scales.

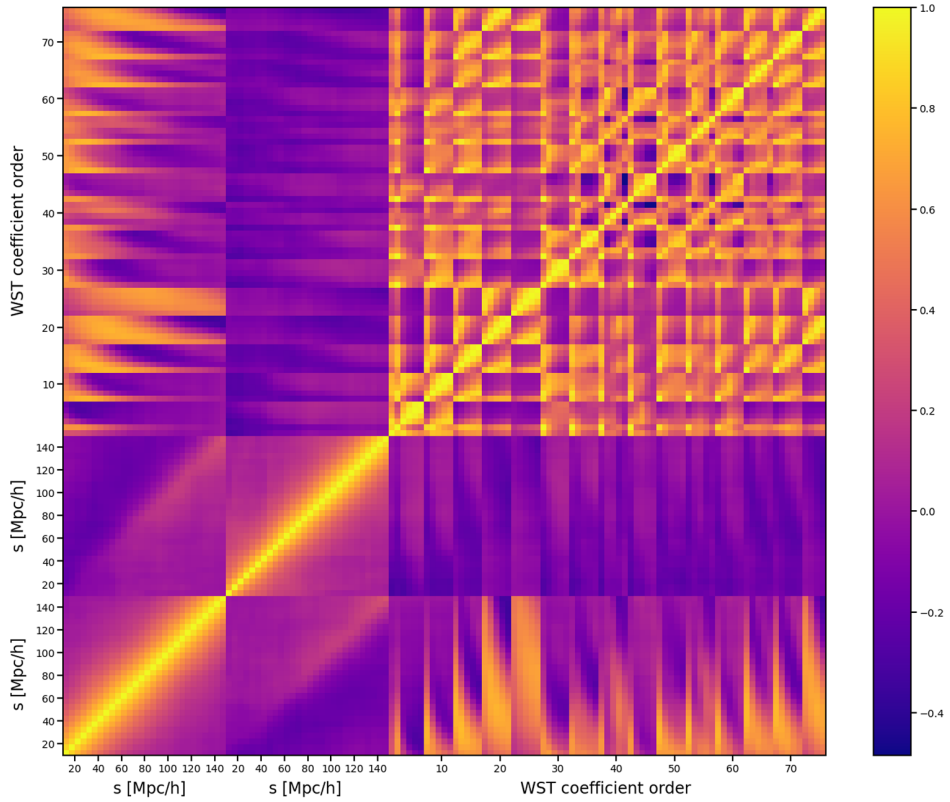

FIG. 5. Correlation matrix of the joint data vector consisting of the multipoles of the 2-point correlation function, $l = \{0, 2\}$, and the 76 WST coefficients used in our analysis, evaluated from the 2048 realizations of the Patchy mocks for the fiducial cosmology. The lower left and upper right subplots coincide with the individual correlation matrices of the two statistics, respectively, while the rest corresponds to the cross-correlations between them. 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 2×2 blocks on the lower left corner correspond to the auto- and cross- correlations of ξ_0 and ξ_2, from bottom to top and from left to right, respectively.

There are no significant correlations with the quadrupole of the correlation function, on the other hand.

The final missing piece needed to evaluate the likelihood (19) for a given point in the target parameter space is the theoretical dependence of the summary statistic as a function of the $8 + 7 = 15$ cosmological + HOD parameters, $\mathbf{X}_t(\theta)$, which we model using the emulator we trained (as explained in Sec. IV E), which allows us to obtain accurate predictions in a fraction of a second.

Combining all of the above ingredients into our model, we sample the likelihood from Eq. (19) using the Markov Chain Monte Carlo (MCMC) sampler EMCEE [139],[11] so as to perform the posterior analysis of the CMASS dataset. Even though our original forward model spans a 15-dimensional parameter space, as explained in Secs. IV A and IV B, our main focus is to obtain constraints on ΛCDM, so we fix $w_0 = -1$, $w_a = 0$, $a_{\rm run} = 0$, $N_{\rm eff} = 3.046$ (i.e. to their ΛCDM values) and define our baseline analysis to constrain the following $4 + 7 = 11$ cosmological + HOD ΛCDM parameters: $\theta = \{\omega_b, \omega_c, \sigma_8, n_s, \log M_{\rm cut}, \log M_1, \sigma, \kappa, \alpha, \alpha_c, \alpha_s\}$. We also obtain constraints on extensions to ΛCDM, for which our analysis will constrain the full 15-d parameter space consisting of 8 cosmological parameters, $\theta = \{\omega_b, \omega_c, \sigma_8, n_s, w_0, w_a, a_{\rm run}, N_{\rm eff}\}$+ the same 7 HOD nuissance parameters as above. We use flat priors bounded by the limits of the AbacusSummit simulations and the HOD training set, both of which are showed in Table I. For parameter ω_b, our baseline run is actually performed with a Gaussian prior:

$$\omega_b = 0.02268 \pm 0.00038, \tag{24}$$

as determined from big bang nucleosynthesis (BBN) measurements, which is a choice commonly adopted in analyses of BOSS data [121–124]. Finally, to confirm the sufficient convergence of our chains, we make sure that the mean value of the acceptance fraction falls within the reasonable range of values, 0.2–0.5, and that the mean integrated autocorrelation time is at least 2 orders of magnitude lower than the total number of steps used, as suggested in Ref. [139]. Lastly, we make use of 8000 walkers, which are initialized in a tight ball around the *Planck* 2018 values.

[11]Publicly available at https://emcee.readthedocs.io/en/stable/.

1. Patchy mocks

The covariance matrix of an estimator can be usually evaluated either using an analytical model under the Gaussian approximation or using simulations performed for multiple realizations at a given cosmology, through Eq. (21). Simulation-based analyses typically take advantage of covariance mocks, such as the 2048 realizations of the publicly available Multidark-Patchy mocks[12] [138,140], hereafter referred to as Patchy mocks. We will use this collection for our BOSS analysis.

The main reference simulation for this run [141] evolved 3840^3 dark matter particles on a cubic volume of side 2.5 Gpc/h, using the code GADGET-2 [142], and for a baseline cosmology described by $\{\Omega_b, \Omega_m, n_s, \sigma_8, h\} = \{0.0482, , 0.307, 0.961, 0.829, 0.6778\}$. It was subsequently combined with an approximate perturbation theory-based gravity solver in order to produce mocks for the gravitationally bound halos, which were identified using the bound density maximum halo finder [143]. Finally, the galaxy mocks were created by populating the halos, using the halo abundance matching technique [144] in order to model the galaxy-halo connection. The Patchy mocks were also cut into the realistic survey geometry of BOSS CMASS, for both the NGC and SGC observed parts of the sky, while the systematic effects can be captured through a set of accompanied weights (in analogy to Eq. (14) for the data):

$$w_c(\mathbf{r}) = w_{fc}(\mathbf{r}) w_{veto}(\mathbf{r}). \tag{25}$$

Similar to Eq. (13), the above weighting scheme captures fiber collisions, w_{fc}, and the rest of the associated shortcomings of the dataset through a veto mask, w_{veto}, while the FKP weights are also assigned through the usual Eq. (15). To evaluate the summary statistics from this set of mocks, we repeat the procedure detailed in Sec. IV D for Eq. (13), but with the weighting scheme in Eq. (25), as opposed to the one of Eq. (14) that we used for the data. For this purpose, the Patchy mocks are also accompanied by their own set of randoms containing ∼50× the number of objects in the corresponding actual galaxy mock.

Furthermore, we follow the standard procedure of assuming a cosmology-independent covariance matrix [145,146], and convert the galaxy coordinates of the mocks, RA, DEC, and z, into comoving Cartesian ones using the fiducial cosmology of our forward model, which is the `c000` defined in Sec. IV A. As we also noted in Ref. [80], mixing different ways of modeling the cosmological dependence of the estimator and its covariance matrix is common practice in BOSS analyses (as in, e.g., Refs. [23,94,95,121–123]). We combine two different sets of mocks (AbacusSummit and Patchy) in order to build our final model for the likelihood.

[12]Available at https://data.sdss.org/sas/dr12/boss/lss/dr12_multidark_patchy_mocks/.

It should be pointed out that, even though the Patchy mocks were also partly tested for their accuracy in capturing the 3-point correlation function of CMASS [138,140], in addition to the 2-point function, they have not been tuned for novel summary statistics such as the WST. This fact, combined also with their approximate gravity solver and the assumption of the cosmology-independent covariance matrix may be sources of error in our analysis, that we are working to overcome with the next generation of galaxy mocks designed to match the requirements for DESI analysis (see such an example in Ref. [147]).

B. Validation

In the previous section we described the detailed steps to perform a likelihood analysis using our emulator for the cosmological dependence of the WST coefficients. Before proceeding to analyze the actual CMASS dataset, we first test our pipeline to ensure its accuracy in inferring (known) cosmological parameters from simulated data vectors.

1. Abacus hold-out mock tests

In order to test the accuracy of our inference pipeline, we begin by randomly selecting 10 HOD configurations centered around the best-fit values from Ref. [103], for each one of the hold-out `c001-004` AbacusSummit cosmologies. We repeat all previously explained steps to produce synthetic WST data vectors from each test mock, which are then fed into our likelihood analysis pipeline to constrain the cosmological parameters of our ΛCDM baseline case. The corresponding marginalized posterior distributions obtained on the 4 ΛCDM cosmological parameters of the baseline analysis are then shown in Fig. 6, in which we see that we are able to recover the true values within 1-σ levels of accuracy, for all cases. We note that we do not show the contours for the wCDM test cosmology `c002` in Fig. 6, for brevity, but the recovery is successful in this case as well.

The hold-out cosmologies used for the above tests correspond to the same initial fixed phase, ph000, of the AbacusSummit simulations as the mocks of our training set, and as a result do not allow us to detect potential biases introduced by this approximation. To check for this, we also attempt to perform parameter inference from the data vector obtained by averaging over the 24 additional phases, ph001-024, that are available for the fiducial `c000` cosmology. As we also show in Fig. 6, our phase correction scheme (22) is found sufficient to recover an accurate cosmology from a different phase. We add that we confirmed the recovery was also successful when we used these 24 phases individually, as the mock data vector.

Overall, the above tests confirm that our WST emulator, in combination with the error correction strategies (20) and (22), is successful in inferring the parameters of the AbacusSummit simulations within 1-σ levels of accuracy, over a wide range of cosmologies. We should also note, at this

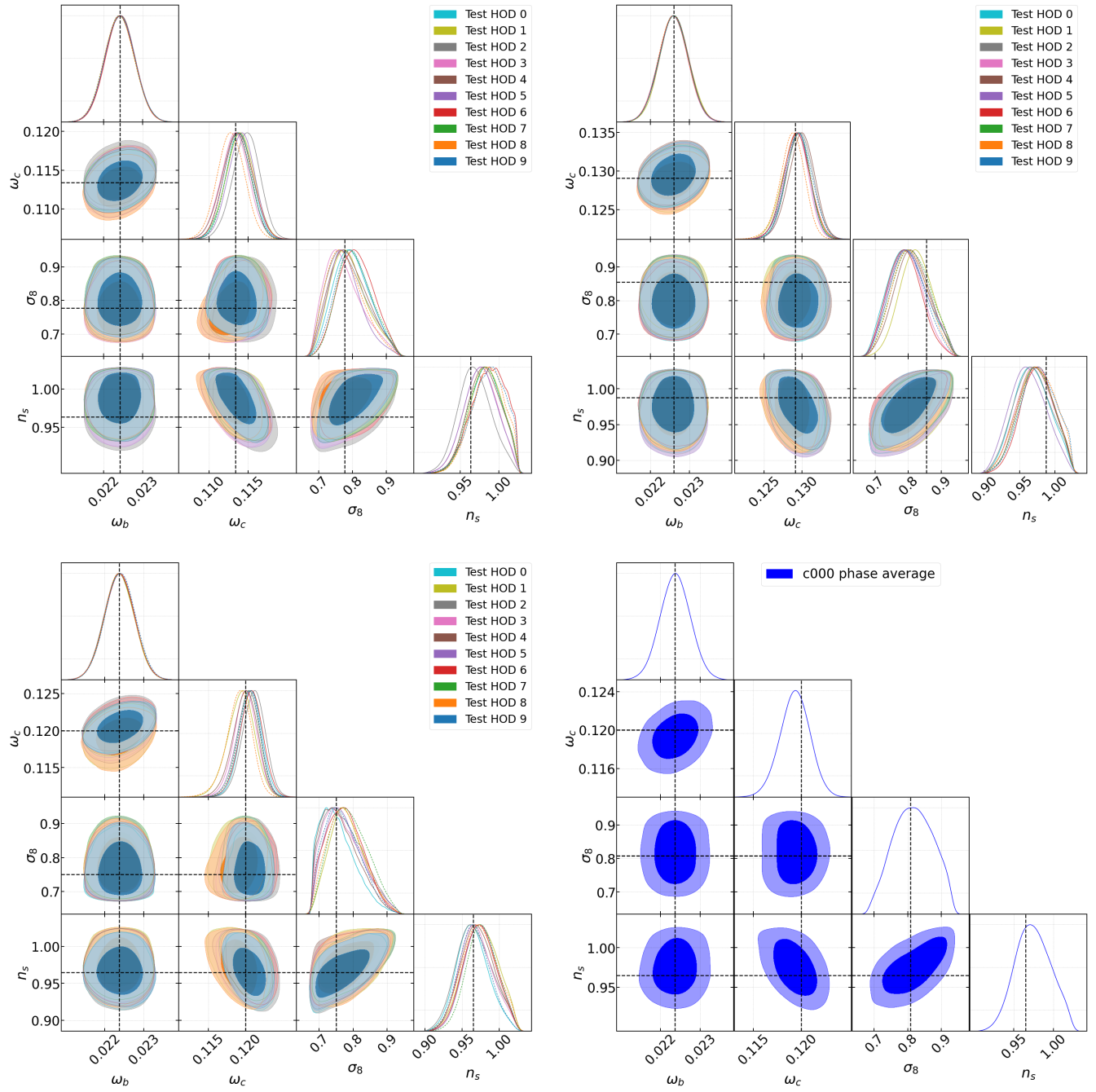

FIG. 6. ΛCDM recovery tests using our WST emulator to analyze 10 HOD configurations of the c001 (upper left), c003 (upper right) and c004 (lower left) hold-out cosmologies of our test set. We also show the marginalized 1-σ and 2-σ posteriors obtained by analyzing the mean data vector of the 24 additional realizations available for the fiducial c000 cosmology (lower right). The horizontal and vertical black dashed lines indicate the true values of the cosmological parameters in each case.

point, that we confirmed the same to be true for the corresponding emulator for the multipoles of the galaxy correlation function.

2. *Uchuu* mock tests

Even though our simulation-based model was found to be successful at recovering the true cosmological parameters over a broad range of tests, as reported in the previous Sec. V B, the corresponding hold-out mocks that we used were produced from the same set of the AbacusSummit simulations, using, more importantly, the same assumptions for the galaxy-halo connection through the specific HOD model we adopted (described in Sec. IV B). As a result, before our pipeline can be trusted for a reliable interpretation of the actual observations, it should be first

tested against an independent simulation with different gravity and halo codes, with a different strategy for the population of dark matter halos with galaxies.

To achieve this goal, we additionally make use of the Uchuu simulations [148–152], which were run using the GreeM N-body code [153]. It evolved 2.1×10^{12} dark matter particles, in a simulation volume $(2\ \mathrm{Gpc/h})^3$, which matches the one of Abacus, and it is large enough to fit the entire footprint of BOSS. Their underlying cosmology corresponds to the following values: $\Omega_m = 0.3089$, $\Omega_b = 0.0486$, $h = 0.6774$, $\sigma_8 = 0.8159$, and $n_s = 0.9667$, while dark matter halos were identified with the Rockstar halo finder [154], in contrast to Abacus's CompaSO.

More crucially, the corresponding galaxy mocks were produced with UniverseMachine [UM; [155]], a model that is considerably more sophisticated than the HOD. UM is an empirical galaxy–halo connection model that predicts galaxy star formation rates from halo mass and halo assembly histories. It is a flexible framework that models the full evolution histories of galaxies anchored on dark matter halo merger trees from cosmological simulations, and it is simultaneously constrained by observed galaxy stellar mass functions, UV luminosity functions, quenched fractions, cosmic star formation history, and galaxy clustering over a wide range of galaxy mass and redshifts (up to $z \sim 8$). We refer the readers to [151] for detailed descriptions of the mock. It is also worth highlighting that UM naturally includes a motivated yet flexible prescription of galaxy assembly bias, as the galaxy properties are directly computed from the halo merger trees. Thus, this test also checks against potential systematic biases due to galaxy assembly bias.

For the covariance matrix needed for the Uchuu likelihood analysis we use the same suite of Patchy mocks described in Sec. V A 1, since both types of simulations are tuned to match the clustering properties of the BOSS CMASS sample, with a same volume and number density and a similar Planck-like cosmology.

In Fig. 7, we plot the marginalized constraints obtained on the 4 ΛCDM parameters after analyzing the Uchuu mock using the multipoles of the galaxy correlation function, the WST coefficients and a joint combination of both. As in the previous case of the Abacus hold-out tests, we find that the true values always lie within 1-σ away from the mean, for all 3 cases. We note that n_s is prior-dominated in the case of the 2-point correlation function, as the contour hits the upper prior bound of the AbacusSummit grid.[13] This is not the case, however, for the WST and the joint combination, which are the main focus of this analysis, despite the significantly tighter 1-σ errors they predict. These results confirm our ability to trust that our forward model can recover unbiased cosmological constraints which are robust against the various assumptions made by the simulations used for its training.

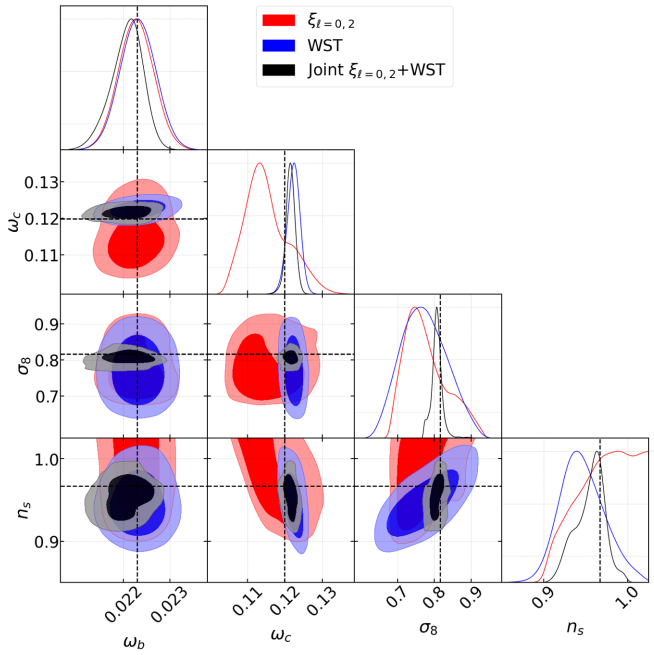

FIG. 7. Recovery test using the Uchuu galaxy mock for the ΛCDM cosmological parameters obtained using the monopole and quadrupole of the galaxy correlation function (red), the WST coefficients (blue) and their joint combination (black). The horizontal and vertical black dashed lines indicate the true values of the cosmological parameters.

VI. RESULTS

Having validated our pipeline against a series of internal and external mock recovery tests, described in Sec. V B, we now proceed to use it in order to analyze the BOSS CMASS dataset. Specifically, in Fig. 8 we plot the 2-dimensional marginalized posterior probability distributions of the 4 ΛCDM parameters of our baseline CMASS analysis, as they were obtained using the multipoles of the galaxy 2-point correlation function, the WST coefficients and their joint combination. We also show the constraints on the dimensionless Hubble constant, h, that we obtain by treating it as a *derived* parameter from our MCMC chains, resulting from the fixed value of the comoving angular size of the sound horizon at last scattering, $100\theta_\star = 1.041533$, imposed in the AbacusSummit simulations. We note that, even though this parameter is very well-constrained by the *Planck* satellite [102], this choice implies that h is not varied independently in our inference, so the corresponding result should be interpreted with caution. In addition, the mean and 1σ error values obtained on the cosmological parameters (marginalized over HOD) are listed in Table II, while the corresponding constraints on the HOD nuisance parameters are presented in Appendix B.

[13] A similar finding was recently reported by Ref. [52].

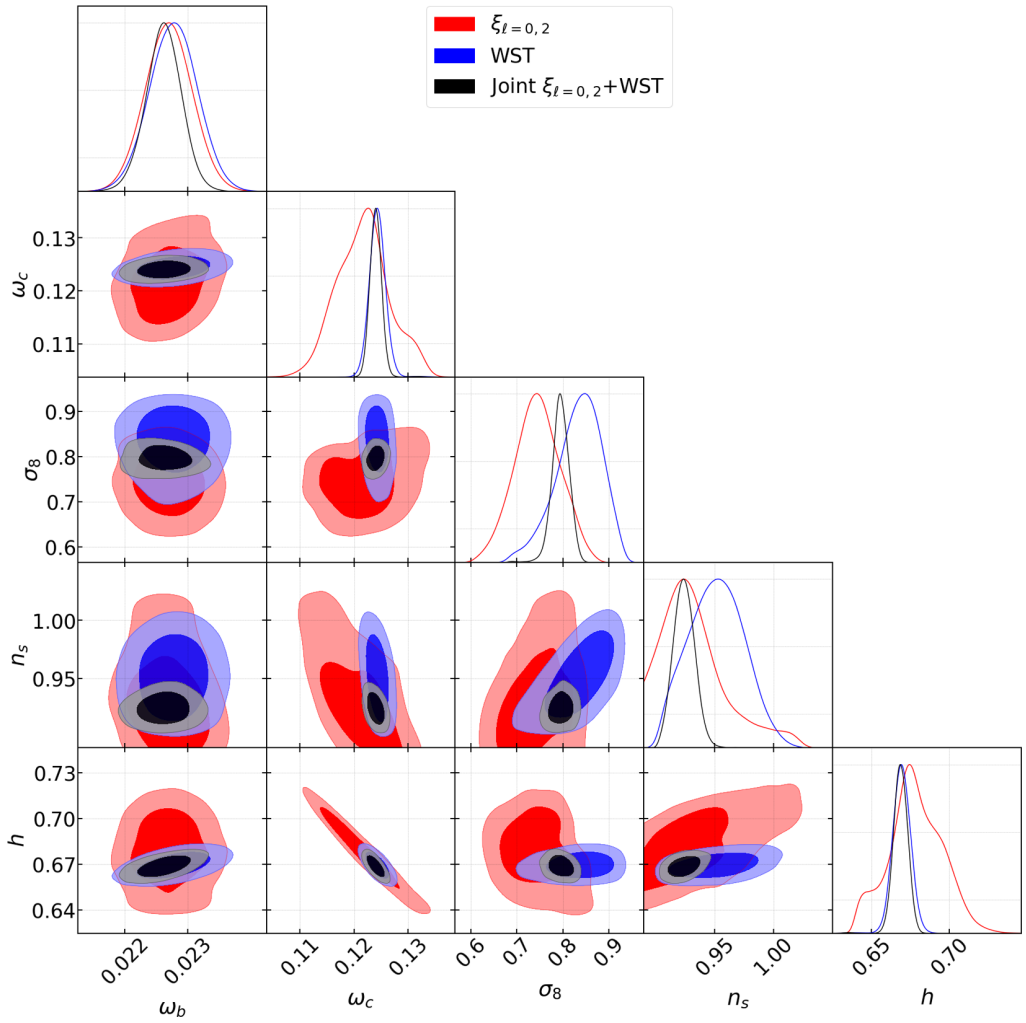

FIG. 8. Marginalized constraints on the ΛCDM cosmological parameters obtained using the monopole and quadrupole of the galaxy correlation function (red), the WST coefficients (blue) and their joint combination (black) in order to analyze the BOSS CMASS observations. The results shown above were obtained after imposing a BBN Gaussian prior on the value of $\omega_b = 0.02268 \pm 0.00038$.

We begin with the standard analysis using the galaxy correlation function, the results of which are broadly consistent with *Planck* 2018 [102]. Even though the mean values obtained for n_s and σ_8 are somewhat lower than the ones of *Planck*, the magnitude of these differences is not statistically significant ($\sim 1\sigma$), unlike the results of some previous BOSS analyses (e.g. [23,103,124]).

Moving on to discuss the results of the WST reanalysis, we first 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-σ values obtained from the correlation function. We do, however, notice different degeneracy directions exhibited by the WST contours projected on the various individual 2-d parameter planes, the importance of which will become apparent below. More importantly, the 1-σ errors obtained on parameters ω_c and n_s are found to be 4.2 and 1.6× tighter than the corresponding predictions from the correlation function, as seen in Table II. The dimensionless Hubble constant is consistent with Planck in this case as well, with an error that is 3.7× tighter than from the correlation function, tracing the respective results for ω_c, on which it depends through the fixed $\theta_\star$. We note again that this finding should be interpreted with caution. if it were not for the strong prior on $\theta_\star$ in our model, the coarse logarithmic binning used by our wavelets would likely not be able to fully capture the BAO information, resulting in a less accurate determination of h. Last but not least, we do not find any noticeable improvement (with respect to the 2-point function) in our ability to constrain σ_8, while the mean value predicted by the WST analysis is also consistent with Planck. Even though counterintuitive, at face value, this result is attributed to the inclusion of the residual emulator error, $C_{\rm emu}$, in Eq. (20), a fact that we have tested and confirmed, as shown in Appendix C. In particular, we find that, even though the intrinsic errors predicted by the WST in the limit of zero emulation error are substantially tighter for all parameters, our actual error

TABLE II. Best-fit values, mean values and 68% confidence intervals for all cosmological parameters resulting from the likelihood analysis of the 2-point correlation function multipoles (left), the WST coefficients (middle) and a joint analysis of the two (right). The mean values are presented in the format $\text{mean}^{+1\sigma}_{-1\sigma}$, after marginalization over all HOD parameters.

	2-point c.f.		WST		Joint 2-point c.f. + WST	
	Best-fit	Mean $\pm\,\sigma$	Best-fit	Mean $\pm\,\sigma$	Best-fit	Mean $\pm\,\sigma$
ω_b	0.02261	$0.02270^{+0.00037}_{-0.00037}$	0.02274	$0.02277^{+0.00038}_{-0.00038}$	0.0225	$0.02262^{+0.00029}_{-0.00029}$
ω_c	0.1201	$0.1222^{+0.0040}_{-0.0063}$	0.1239	$0.1244^{+0.0015}_{-0.0015}$	0.1238	$0.1241^{+0.0011}_{-0.0011}$
n_s	0.925	$0.922^{+0.037}_{-0.037}$	0.961	$0.951^{+0.023}_{-0.023}$	0.927	$0.924^{+0.01}_{-0.01}$
σ_8	0.742	$0.746^{+0.051}_{-0.051}$	0.860	$0.834^{+0.058}_{-0.039}$	0.793	$0.795^{+0.019}_{-0.019}$
h	0.677	$0.677^{+0.022}_{-0.015}$	0.67	$0.669^{+0.0059}_{-0.0059}$	0.668	$0.669^{+0.0049}_{-0.0049}$

budget is also much larger for the WST, such that the addition of $C_{\rm emu}$ in Eq. (20) partially or completely (in the case of σ_8) masks the net improvements. We nevertheless choose to include this term, in order to ensure a reliable and robust analysis.

We have already seen that, despite the fact that the WST and correlation function contours are consistent with each other at the 1σ level, they exhibit different degeneracy orientations. This is not as surprising if we consider that the two statistics do not capture the exact same information. Indeed, we remind that, even at the lowest (1st) order, the WST raises the modulus of the input galaxy density field to the power $q = 0.8$, closer to the properties of other higher-order statistics such as the marked power spectrum, as we also found in Ref. [79]. The localized solid harmonic wavelet (4) is also different than the Fourier kernel of the power spectrum/correlation function, with all the additional known benefits associated with this choice [70–73]. As a consequence, analyzing the data with the joint combination of the two statistics allows us to break degeneracies and further improve upon the results obtained from each individual analysis, as we can see in Fig. 8 and Table II. In particular, the 1σ error obtained on σ_8, which previously did not improve by a WST application alone, now shrinks by a factor of 2.5, while the corresponding constraints on the rest of the parameters are further tightened by a factor of 3–6 compared to the 2-point correlation function and by a factor of 1.4–2.5 compared to the WST-only results. Overall, the joint analysis allows us to constrain the parameters ω_c, σ_8, n_s, and h with 0.9%, 2.3%, 1% and 0.7% levels of accuracy, respectively. This result, which can be considered to be the main one of our work, highlights the value held in a complementary analysis employing both the WST coefficients and the standard correlation function.

In addition to the above parameters, and in order to align our analysis with a standard practice adopted by many RSD studies in the literature, we further quote results on the product $f(z)\sigma_8(z)$, with $f(z) = \frac{d\ln D(a)}{da}$ and $D(a)$ being the linear growth rate and growth factor, respectively. This is also a derived parameter that we obtain from the samples in our chains. In particular, at the effective redshift, $z_{\rm eff} = 0.515$, of our sample, the joint analysis gives:

$$f\sigma_8(z_{\rm eff} = 0.515) = 0.469 \pm 0.012, \tag{26}$$

which corresponds to a determination at a 2.5% level of accuracy. Furthermore, in Fig. 9 we plot our result together with the corresponding one from *Planck* 2018 [102] and from a selected sample of other recent BOSS reanalyses in the literature (which we will further discuss shortly). Our prediction is consistent with *Planck* well within the 1σ levels, driven by the corresponding consistency in our inferred value of σ_8, and despite our relatively higher value of ω_c. In the context of lensing studies, this can be alternatively examined in terms of the parameter combination $S_8 = \sqrt{(\Omega_m/0.30)}\sigma_8$, for which we get

$$S_8 = 0.833 \pm 0.023, \tag{27}$$

in almost perfect agreement with the fiducial *Planck* result, $S_8 = 0.832 \pm 0.013$.

As far as the values of ω_c and n_s are concerned, they are found to be statistically higher and lower, respectively, relative to the Planck result, driven by the very tight constraints of our joint analysis. It is interesting to notice that a similar trend has been found in some recent large-scale BOSS analyses, when n_s is left free [23]. The magnitude of the tension was smaller in these studies, however, due to the larger error bars produced by such perturbation theory-based models. Reference [103] also found a preference for a lower n_s at the 1.5σ level.

We also note that, even though all above results were produced assuming a tight BBN prior (24) on ω_b, we found that our joint analysis is actually able to constrain this value reasonably well even with a flat, uninformative prior, as shown in Appendix D.

Furthermore, in the left sub-panels of Table II we list the best-fit values obtained for each one of the 3 types of analyses considered in this work, which are always found to

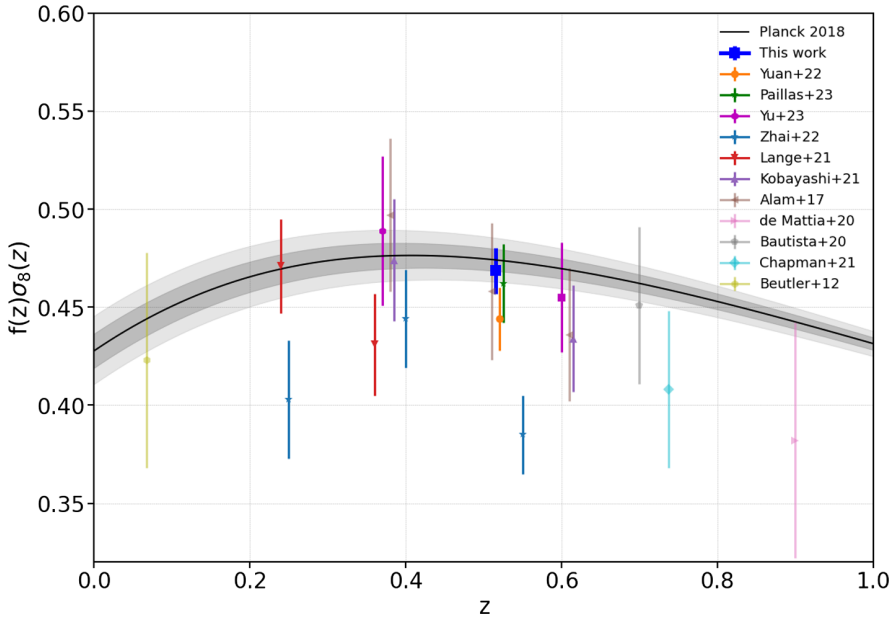

FIG. 9. Marginalized constraints on the structure growth rate, $f(z)\sigma_8(z)$, of our joint analysis in blue alongside other clustering constraints in the literature. We show the *Planck* 2018 [102] CMB constraints in black, with the corresponding 68% and 95% limits in shaded bands, together with the results from two other recent Abacus-based CMASS reanalyses using the small-scale 2-point correlation function [103] and the density split clustering statistic [51,52]. Additionally, we show clustering constraints from BOSS LOWZ small-scale RSD [97], BOSS full-shape power spectrum [135], BOSS large-scale RSD + BAO [156], BOSS small-scale RSD [157] eBOSS small-scale RSD [158], eBOSS large-scale RSD + BAO [159,160], BOSS DR12 large-scale power spectrum [161] and the 6dF Galaxy Survey [162].

lie within a standard deviation away from the corresponding means. To assess the goodness-of-fit, we also evaluate the χ^2 per degrees of freedom (d.o.f.), $\chi^2_\nu \equiv \chi^2/\text{d.o.f.}$, which is found to be equal to 1.11, 1.36 and 1.37 for the correlation function, the WST and the joint analysis, respectively. The result for the 2-point function is very similar to the one reported by the Abacus-based small-scale CMASS analysis of [103]. Even though the value for the WST is a bit higher, it is still reasonable and within the same range and/or lower than the corresponding results reported by recent analyses using other higher-order statistics, such as k-nearest neighbors [163] and density-split statistics [51]. The goodness of the fit is also visually evident in Fig. 10, in which we plot the best-fit prediction for the WST together with the corresponding CMASS measurement.

We should comment, at this point, on how our new results compare with the ones of our previous WST BOSS analysis [80], which relied on a simple Taylor expansion approximation to model the cosmological dependence of the WST coefficients. Starting with the 1-σ errors, the main difference for parameters σ_8 and n_s is caused by the inclusion of the WST emulator error, C_{emu}, in Eq. (20), as we also pointed out above and show in Appendix C. In fact, we note that, if we omit this contribution, the WST errors on σ_8 and n_s become 0.027 and 0.02, respectively, which are not that far off from what we previously reported, as seen in Table II of Ref. [80]. Even after neglecting the emulator error, the constraint on ω_c was still found to be ∼2.5× tighter in Ref. [80], a fact that is most likely attributed to the simplified Taylor expansion approximation. This fact is also most likely responsible for the relatively low value of σ_8 that we reported in that work.

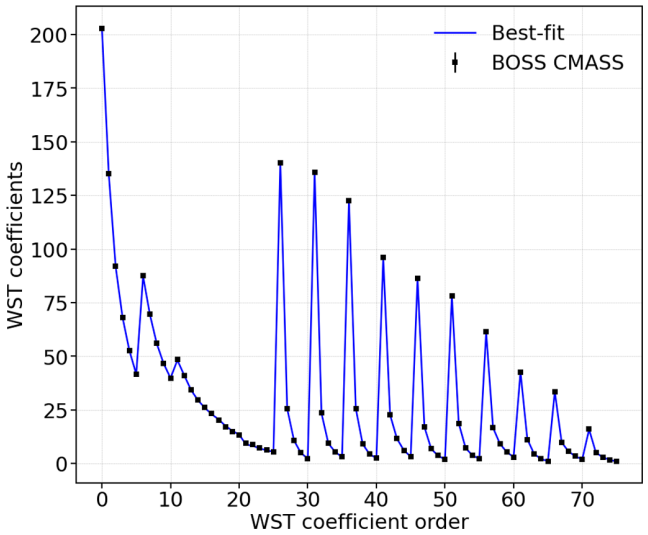

FIG. 10. All 76 WST coefficients evaluated from the BOSS CMASS dataset (black circles) plotted together with the best-fit prediction obtained from our likelihood analysis (solid blue line). The WST coefficients populate the data vector in order of increasing values of the j_1 and l_1 indices, with the l_1 index varied faster.

Finally, we briefly discuss how our work compares against other analyses of BOSS clustering in the literature, starting with the two other recent applications of the AbacusSummit. Reference [103] built an Abacus-based emulator of the anisotropic 2-d correlation function to analyze the CMASS dataset, while Refs. [51,52] worked with the density-split clustering statistic. Given that they relied on the same suite of simulations for their forward model, these applications share the same cosmology grid and priors as our analysis (including the fixed value of sound horizon θ_*). All three analyses also used the same HOD framework. However, there are several key differences between the above works and ours, that need to be pointed out: given the unique sensitivity of the WST to the survey geometry, through the successive wavelet convolutions in Eq. (2), we trained our emulator using the cut-sky mocks matching the exact CMASS footprint, rather than the original cubic boxes used by [51,52,103]. For similar reasons, we worked with a flattened density profile, $n(z)$, and in a slightly narrower redshift cut, as we showed in Fig. 1. Furthermore, both of the other works included smaller scales down to 1 Mpc/h and thus accounted for the necessary effects of assembly bias in their HOD parametrization, which we neglected given that our analysis stopped at a minimum

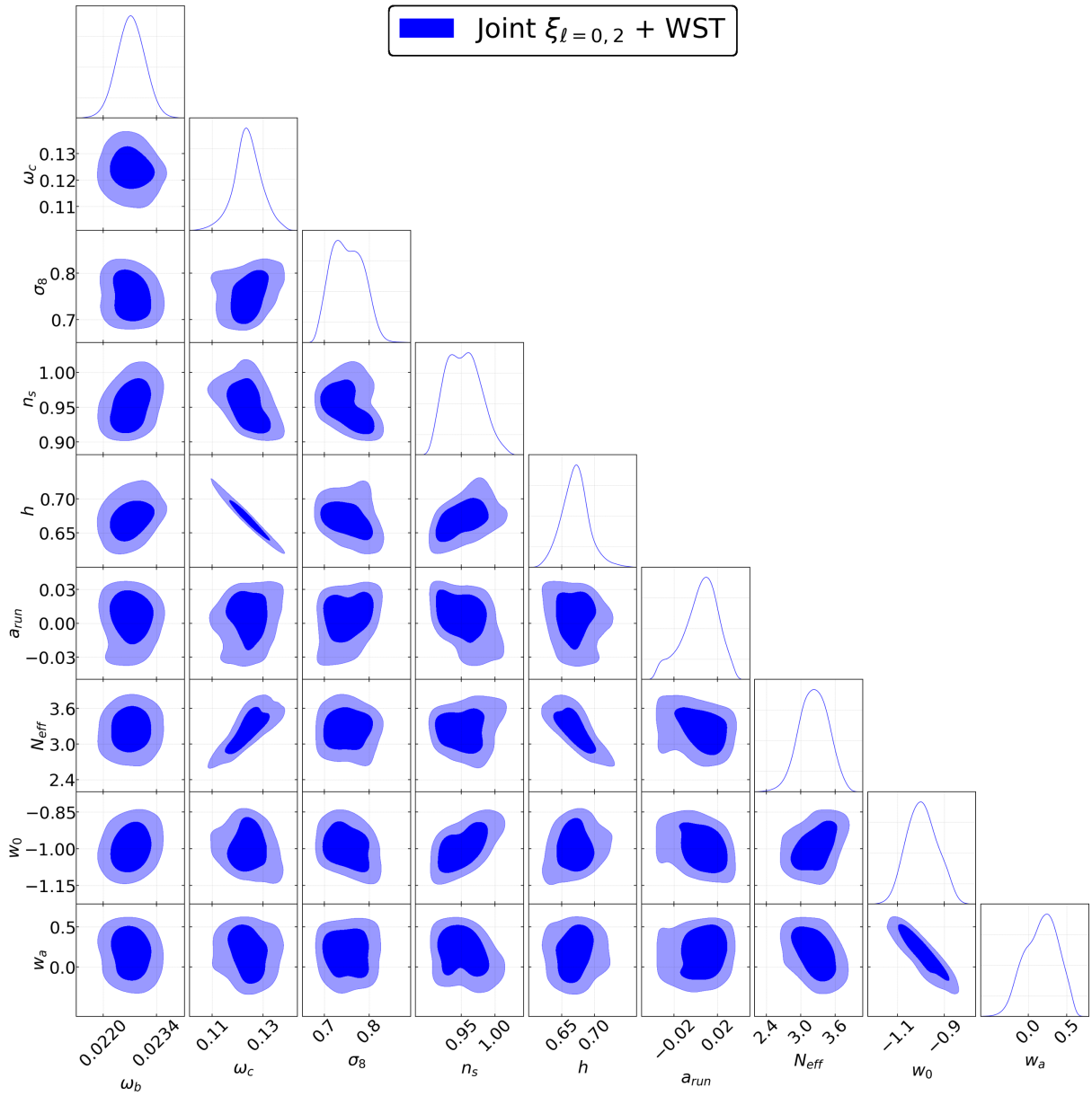

FIG. 11. Marginalized constraints on extensions to ΛCDM obtained using the joint WST + correlation function combination in order to analyze the BOSS CMASS observations. The results shown above were obtained after imposing a BBN Gaussian prior on the value of $\omega_b = 0.02268 \pm 0.00038$.

scale of ∼10 Mpc/h. Reference [103] also used Jackknife resampling in order to compute the covariance matrix (as opposed to our Patchy mocks) and included only small scales, $< \times 30$ Mpc/h, in their analysis, while Ref. [51] only analyzed the NGC part of the BOSS survey. Due to all the above differences, a direct "apples to apples" comparison is still hard. Nevertheless, we notice the relative 1σ consistency between our results for σ_8 and $f\sigma_8$ and those of [51] (and Planck), as also seen in Fig. 9. The analysis of Ref. [103], on the other hand, found a relatively lower value of the clustering amplitude, which, combined with their tighter errors, leads to a disagreement at the level of 1.5σ. In addition to the previously mentioned differences, other analysis choices that might be driving this difference are the use of Gaussian priors by Ref. [103] and the fact that their emulator error was evaluated drawing from from the posterior, rather than the prior. In preparation for the analysis of the next stage of spectroscopic observations, we plan to revisit these comparisons through a commonly adopted set of uniform analysis choices.

A plethora of other studies in the literature have analyzed the BOSS and extended BOSS (eBOSS) [164] observations, including, but not limited to, the ones plotted in Fig. 9 alongside our result. All of them used the standard 2-point correlation function or the power spectrum, and can be grouped into large-scale [156,159–162] and small-scale studies [97,135,157,158]. Despite the large variance in the modeling and analysis choices among the members of this list, we notice that our analysis joins the ones that are statistically consistent with the Planck curve, including the official BOSS result [156]. On the other hand, a number of studies is found to systematically underpredict the growth rate, related to the known LSS tension that has emerged in the last few years. Given that the true origin of this discrepancy is not yet known, we hope that novel techniques such as the WST will help shed light on this issue. We highlight that our constraint is found to be the tightest reported among all these studies. Similar considerations apply for the comparison to other BOSS analyses, e.g., [23,121–125,165].

A. Constraints on ΛCDM extensions

Our main focus for the present analysis has been to obtain constraints on ΛCDM. However, as we explained in Sec. IV A, our emulator was originally trained on the AbacusSummit cosmology grid, which also includes 4 additional parameters describing extensions to ΛCDM: $\{a_{\rm run}, N_{\rm eff}, w_0, wa\}$. As a result, and in order to further explore the constraining capabilities of the WST, here we briefly present constraints from the base joint WST + correlation function likelihood analysis on all 8 cosmological parameters (marginalized over the 7 HOD parameters), shown in Fig. 11 and Table III. We find that our analysis is able to clearly constrain all parameters simultaneously, without any signs of statistically significant deviations away from the known ΛCDM limits, $w_0 = -1$, $w_a = 0$, $a_{\rm run} = 0$, $N_{\rm eff} = 3.046$. Given the increased number of parameters in this case, it is not surprising that the constraints on the ΛCDM parameters are looser compared to the corresponding values found in the base analysis. As a consequence of the same fact, the previously reported tensions for ω_c and n_s are alleviated in this case, and all ΛCDM parameters are found to be consistent with the *Planck* 2018 results [102] (and also with the ones of the base analysis). The reduced χ^2/d.o.f is found to be $\chi^2_\nu = 1.31$, confirming that the fit is equally good as the one of the joint base analysis.

TABLE III. Best-fit values, mean values and 68% confidence intervals for all cosmological parameters resulting from the joint WST + correlation function likelihood analysis in the case of the extended cosmological scenario. The mean values are presented in the format $\text{mean}^{+1\sigma}_{-1\sigma}$, after marginalization over all HOD parameters.

	Joint 2-point c.f. + WST	
	Best-fit	Mean $\pm\,\sigma$
ω_b	0.02280	$0.02273^{+0.00036}_{-0.00036}$
ω_c	0.1227	$0.1239^{+0.0056}_{-0.0056}$
σ_8	0.748	$0.751^{+0.034}_{-0.040}$
n_s	0.928	$0.953^{+0.022}_{-0.030}$
h	0.675	$0.671^{+0.021}_{-0.021}$
$a_{\rm run}$	0.002	$0.004^{+0.019}_{-0.012}$
$N_{\rm eff}$	3.048	$3.23^{+0.26}_{-0.26}$
w_0	−1.039	$-0.995^{+0.061}_{-0.073}$
w_a	0.29	$0.17^{+0.24}_{-0.21}$

We note that our pipeline has been more thoroughly tested for ΛCDM applications, and these results are exploratory in nature, while we reserve a more detailed WST application to extended scenarios for future work. Nevertheless, they serve as an additional example that showcases the promise held in the use of the WST in the context of parameter inference applications.

VII. CONCLUSIONS

In this work, we perform a thorough reanalysis of the BOSS CMASS DR12 dataset, using a simulation-based emulator for the wavelet scattering transform, a novel statistic that promises to capture non-Gaussian information in a clustered field by subjecting it to a series of successive wavelet-convolutions.

In our series of previous works [79,80], we laid the foundation for a WST application to spectroscopic galaxy data, including the methodology to capture all necessary associated layers of realism to achieve this task, such as the effects of nontrivial survey geometry, the shortcomings of the dataset through a set of systematic weights or the Alcock-Paczynski effect. However, in order to reduce the

related computational cost, in Ref. [80] we used a linear Taylor expansion to approximate the cosmological dependence of the WST estimator.

Having the full suite of the state-of-the-art AbacusSummit simulations at our disposal, we now revisit our previous analysis after constructing an accurate neural net-based emulator for the cosmological dependence of the WST coefficients. Our forward model is trained using a total of 151,474 mocks that span a 15-dimensional parameter space, capturing variations in 8 cosmological parameters and 7 Halo Occupation Distribution (HOD) nuisance parameters to model the galaxy-halo connection. We repeat these steps to create a corresponding emulator for the standard multipoles of the galaxy 2-point correlation function, which serves as our benchmark to evaluate the performance of the WST.

In order to ensure that our likelihood analysis pipeline achieves the necessary levels of accuracy for a reliable and robust cosmological analysis, we subject it to a series of internal and external parameter recovery tests. For the internal tests, we use 40 hold-out mocks that span a broad range of cosmological parameters within our training set. We then test and confirm that we can accurately infer the parameters of an external simulation, that used different assumptions to capture the complicated physics of galaxy formation.

After confirming the accuracy of our forward model, we use it to reanalyze the BOSS CMASS DR12 dataset in the redshift range $0.4613 < z < 0.5692$, in order to constrain the ΛCDM parameters using the WST coefficients and the multipoles of the galaxy correlation function. We find that a joint analysis using the WST and the correlation function allows us to constrain the ΛCDM parameters with 1σ errors that are tighter by a factor of 2.5–6, compared to the 2-point correlation function, and by a factor of 1.4–2.5 compared to the WST-only results. This corresponds to a competitive 0.9%, 2.3%, 1% and 0.7% level of determination for parameters ω_c, σ_8, n_s, and h, respectively. Furthermore, the joint analysis allows us to obtain a tight 2.5% constraint on the parameter combination $f(z)\sigma_8(z)$, in agreement with the 2018 results of the *Planck* satellite. We discuss how our new results compare against our previous analysis and prior ones in the literature, reaffirming the constraining power of the WST.

We also obtained constraints on extended cosmological scenarios, parametrized through 4 additional parameters, $\{a_{\rm run}, N_{\rm eff}, w_0, wa\}$, finding no statistically significant deviations from the ΛCDM limit.

Our emulator for the cosmological dependence of the WST coefficients (and the correlation function) has allowed us to overcome the main limitation behind our previous application [80]. There is, however, room for further improvement in certain components of our forward model, which we plan to achieve in future work. First of all, and as we already pointed out above, the AbacusSummit simulations impose a fixed value of the angular scale $\theta_\star$. Even though this quantity is very well constrained by CMB observations [102], such a prior implies that the Hubble constant is not independently varied in our chains, but can only be obtained as a derived parameter. Given, however, that our framework is flexible enough to be applied to any set of simulated mocks, this limitation can be easily overcome using a different training set. In a similar manner, our use of simulations produced at a fixed redshift, $z = 0.5$, implies that clustering evolution along the CMASS light cone is currently neglected. With the capability to produce Abacus light cones already in place [136,166], we plan to incorporate this effect in future revisions of our pipeline. Furthermore, our current HOD parametrization for the galaxy-halo connection neglected assembly bias, given the more conservative scale-cut we adopted. It would be very interesting, in future work, to explore the full small-scale constraining power of the WST, for which a more general HOD model including assembly bias would be necessary. Such an endeavor will also require a more careful treatment of systematic effects, such as fiber collisions, which we currently corrected using the recipe designed for the standard correlation function analysis (for an example, see Ref. [136]).

The culmination of this series of works opens up an avenue of potentially exciting cosmological applications of the WST, with the advent of the first stage-IV spectroscopic observations by DESI. As we had also pointed out in Ref. [80], the basis of solid harmonic wavelets that we have been using is not optimized for a spectroscopic dataset, as it was designed in the context of isotropic 3-d applications of molecular chemistry. A suitably tailored new basis of wavelets could potentially fully leverage the anisotropic RSD information in the observed galaxy field, by treating a given direction as special [167]. Higher-order statistics, as we have discussed before [45,79], also exhibit tremendous potential for constraining fundamental physics such as massive neutrinos, theories of gravity or primordial non-Gaussianity, through their unique ability to break degeneracies that are present at the power spectrum level. All of these are very interesting avenues that we would like to explore, alongside the first WST application to the first year of DESI data.

Our application serves as a prime example of how novel estimators, such as the wavelet scattering transform, can hopefully allow us to fully exploit the vast amount of information that will be accessed by the next generation of cosmological surveys, giving us the opportunity to potentially revolutionize our fundamental understanding of the universe.

ACKNOWLEDGMENTS

We would like to thank Carolina Cuesta-Lazaro, Daniel Eisenstein, Hector Gil-Marin, Johannes Lange, Enrique Paillas and Peter Taylor for useful discussions over the

course of this work. G. V. acknowledges the support of the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship at the University of Chicago, a Schmidt Futures program. C. D. and G. V. have been partially supported by the Department of Energy (DOE) Grant No. DE-SC0020223. 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 Fısica Teorica UAM/CSIC, and NERSC at the Lawrence Berkeley National Laboratory. That work acknowledges support from the Spanish MICINNs Consolider-Ingenio 2010 Programme under grant 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. We thank Instituto de Astrofisica de Andalucia (IAA-CSIC), Centro de Supercomputacion de Galicia (CESGA) and the Spanish academic and research network (RedIRIS) in Spain for hosting Uchuu DR1 and DR2 in the Skies & Universes site for cosmological simulations. The Uchuu simulations were carried out on Aterui II supercomputer at Center for Computational Astrophysics, CfCA, of National Astronomical Observatory of Japan, and the K computer at the RIKEN Advanced Institute for Computational Science. The Uchuu DR1 and DR2 effort has made use of the skunIAA RedIRIS and skun6IAA computer facilities managed by the IAA-CSIC in Spain (MICINN EU-Feder Grant No. EQC2018-004366-P).

APPENDIX A: GAUSSIANITY OF THE WST LIKELIHOOD

Our posterior analysis in Sec. V has been performed by sampling from a likelihood that we assumed to follow a Gaussian form, as given by Eq. (19). In the standard power spectrum case, and despite the non-Gaussianity of the cosmic density field at late times, this is known to be an accurate approximation thanks to the central limit theorem, when a sufficiently large number of modes contribute to the value evaluated at a given spatial bin. For the WST, in our previous applications [79,80] we also adopted this approximation, motivated by supportive findings in the 2D weak-lensing (WL) applications of Ref. [78]. We now proceed to explicitly test and confirm the validity of this assumption for the present WST application to 3D clustering, using the 2048 realizations of the Patchy mocks for the fiducial cosmology. Following Refs. [49,168], the 2048 realizations will have a χ^2 distribution, given by:

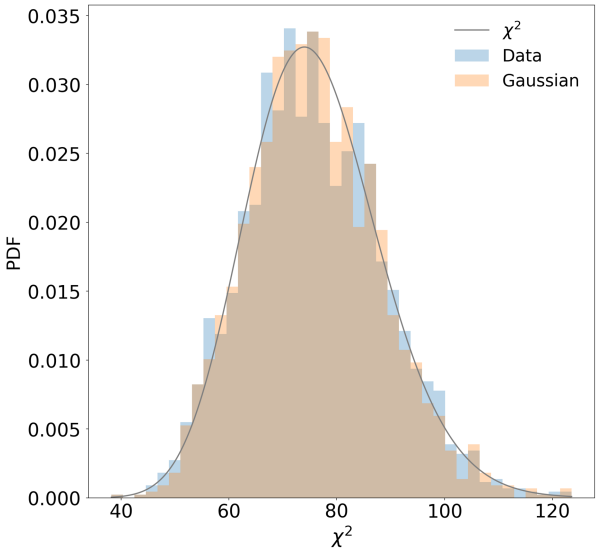

FIG. 12. Probability density function of the χ^2 distribution of the WST coefficients as measured from the 2048 realizations of the Patchy mocks (blue) plotted together with a theoretical χ^2 distribution with $N_{\mathbf{d}} = 76$ degrees of freedom (black line) and a Gaussian distribution with the same mean and covariance (orange). The WST estimator does not exhibit any significant deviations from a Gaussian distribution.

$$\chi_i^2 = [\mathbf{X}_{\mathbf{d}_i} - \bar{\mathbf{X}}_{\mathbf{d}}]^{\mathrm{T}} C^{-1} [\mathbf{X}_{\mathbf{d}_i} - \bar{\mathbf{X}}_{\mathbf{d}}], \tag{A1}$$

where $\mathbf{X}_{\mathbf{d}_i}$ is the prediction for the ith Patchy mock realization, $\bar{\mathbf{X}}_{\mathbf{d}}$ the mean value over the distribution, and C the covariance matrix from Eq. (21).

If the likelihood of a summary statistic is indeed Gaussian, then the probability density function (pdf) from Eq. (A1) should closely track the theoretical χ^2 distribution with degrees of freedom equal to the dimensionality of the data vector (i.e. $N_{\mathbf{d}} = 76$ for our WST implementation). It should also closely match the pdf of samples randomly drawn from a Gaussian distribution with the same mean and covariance as the sample of realizations. This comparison is demonstrated in Fig. 12 for the WST, which is observed to satisfy a high level of consistency between the 3 curves, confirming thus a high degree of Gaussianity for the likelihood of the WST estimator. The equivalent comparison for the 2-point correlation function multipoles is shown in Fig. 13 for reference, which reproduces the known result of Gaussianity of the correlation function. This result for the Gaussianity of the WST is aligned with the one of Ref. [78] in the context of weak lensing and also with the results of Refs. [49,136,168] for other higher-order statistics explored in the literature.

We also note that a quantification of the Gaussianity of various summary statistics was performed in Ref. [169], in which the probability distribution of the WST coefficients evaluated from the simulated 3D matter density field was found to exhibit a certain degree of non-Gaussianity.

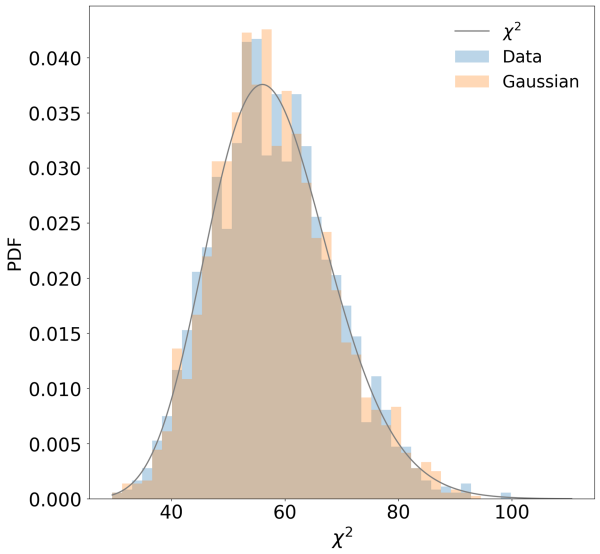

FIG. 13. The same χ^2 analysis as in Fig. 12 is repeated for the multipoles of the 2-point correlation function that we use as the benchmark in our analysis.

However, this work used a different basis of wavelets, that performed a much finer sampling of the spatial domain, which can lead to a breakdown of the central limit theorem. As a result, their findings are not inconsistent with ours.

APPENDIX B: CONSTRAINTS ON HOD PARAMETERS

In this section, we present the constraints obtained on the full set of HOD parameters of our WST and joint WST + correlation function analyses, shown in Fig. 14 and Table IV. We find that the WST alone is capable of constraining all HOD parameters of our model, with only modest additional improvements delivered after the inclusion of the 2-point correlation function. Our analysis hints at a preference for nonzero velocity biases both for the central and also for the satellite galaxies, through the corresponding inferred values for parameters α_c and α_s. Even though the former result is in agreement with the small-scale CMASS reanalysis of Ref. [103], it is interesting that the same work did not find a preference for a satellite bias. We defer a more detailed investigation of this matter to future work, which will extend our analysis to equally small scales.

TABLE IV. Best-fit values, mean values and 68% confidence intervals for the 7 nuisance HOD parameters of our base likelihood analysis using the WST coefficients (left) and the joint analysis of the 2-point correlation function + WST (right). The mean values are presented in the format mean$^{+1\sigma}_{-1\sigma}$.

	WST		Joint 2-point c.f. + WST	
	Best-fit	Mean $\pm\,\sigma$	Best-fit	Mean $\pm\,\sigma$
$\log M_{\rm cut}$	12.681	$12.668^{+0.068}_{-0.068}$	12.608	$12.613^{+0.045}_{-0.060}$
$\log M_1$	13.34	$13.33^{+0.13}_{-0.13}$	13.252	$13.25^{+0.11}_{-0.11}$
$\log \sigma$	−0.783	$-0.823^{+0.11}_{-0.097}$	−0.829	$-0.87^{+0.25}_{-0.25}$
α	0.921	$0.934^{+0.064}_{-0.054}$	0.943	$0.944^{+0.077}_{-0.049}$
κ	1.336	$1.36^{+0.32}_{-0.32}$	1.236	$1.22^{+0.28}_{-0.28}$
$\alpha_{\rm c}$	0.322	$0.34^{+0.17}_{-0.20}$	0.367	$0.32^{+0.16}_{-0.22}$
$\alpha_{\rm s}$	0.306	$0.32^{+0.12}_{-0.11}$	0.411	$0.408^{+0.099}_{-0.049}$

APPENDIX C: WST CONSTRAINTS WITHOUT EMULATOR ERROR

In Sec. V A, we explained how the residual emulator error, $C_{\rm emu}$, was treated as an additional covariance contribution that we added to the overall error budget, through Eq. (20). In order to illustrate the impact of this factor to the WST constraints, and also to better facilitate the comparison with our previous work [80] (which did not account for the emulator error), we repeat the WST analysis using the contribution from the Patchy mocks only [that is, the first term in Eq. (20)] and contrast it against the full result, in Fig. 15. We notice that the inclusion of the emulator error $C_{\rm emu}$ leads to a substantial increase in the 1σ errors for parameters σ_8 and n_s, in particular, with the corresponding impact being much less significant for ω_c. When we neglect this term, on the other hand, and as we also pointed out in the main text, the constraints become much tighter and comparable to the ones of our previous analysis [80] in the case of σ_8 and n_s. We stress that this result should be interpreted with caution, given that the emulator error has not been accounted for. It does serve, nevertheless, as an indication of the intrinsic constraining power of the WST in the limit of zero emulation error. In order for this potential to actually be exploited by the next stage of precise spectroscopic observations, however, higher accuracy emulators and more precise characterization of the emulator error will be necessary. Whether and how these goals can be achieved is a matter of intense study.

APPENDIX D: IMPACT OF PRIORS ON ω_b

Our main analysis used a tight BBN prior on the value of ω_b, from Eq. (24). In this appendix we repeat our joint WST + correlation function analysis using a flat ω_b prior, instead, and demonstrate the comparison between the two results in Fig. 16. Remarkably, we find that the joint analysis is also able to accurately constrain ω_b, as well as the rest of the parameters, using completely uninformative priors. The corresponding increase in the 1σ errors is 90% for ω_b and no more than 10% for the rest of the 3 parameters.

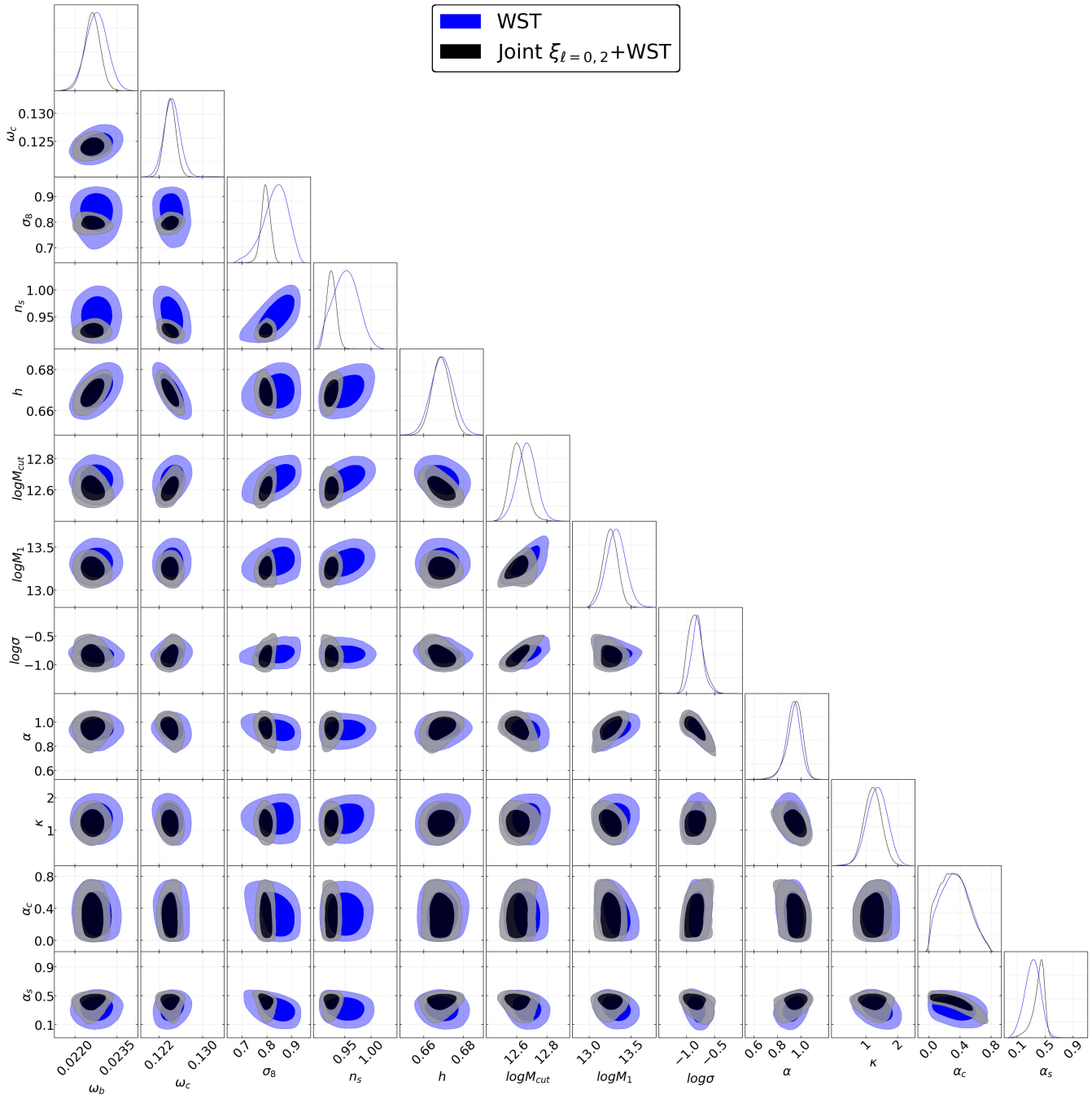

FIG. 14. Marginalized constraints on the full set of cosmological + HOD parameters obtained using the WST coefficients (blue) and the joint combination of WST + correlation function multipoles (black) in order to analyze the BOSS CMASS observations. The results shown above were obtained after imposing a BBN Gaussian prior on the value of $\omega_b = 0.02268 \pm 0.00038$.

APPENDIX E: SENSITIVITY TO SMALL SCALES

In principle, the solid harmonic wavelets that we use in this analysis do not have a finite support neither in real or Fourier space. The Fourier transform of the radial part of Eq. (4) is a Gaussian for $\ell = 0$, that can extend to higher k for $\ell > 0$ values. Even though in practice this can be controlled through a sufficiently conservative combination of the Gaussian width and grid size, as we did in this application, we need to explicitly make sure that our WST analysis does not extract information from smaller scales than originally intended. We confirm this fact through the following test: we first apply a sharp top-hat filter in k-space to our galaxy field and, after going to real space, use this filtered field instead as the input into the WST scattering network (6). This addition imposes a sharp k-space cutoff, which would remove any

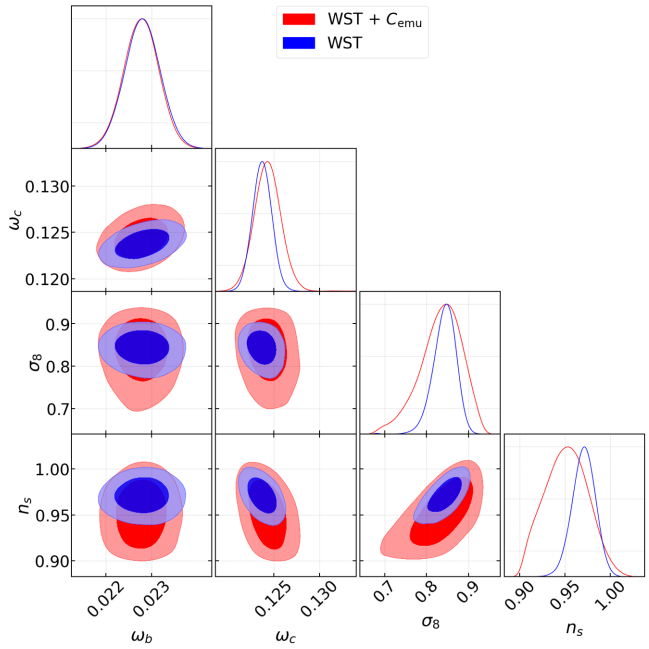

FIG. 15. Marginalized constraints on the ΛCDM cosmological parameters obtained using the WST coefficients without the inclusion of the emulator error, $C_{\rm emu}$, in Eq. (20), shown in the blue contours. The result of the main WST analysis using the full covariance (originally shown in Fig. 8) is also plotted in red, for comparison.

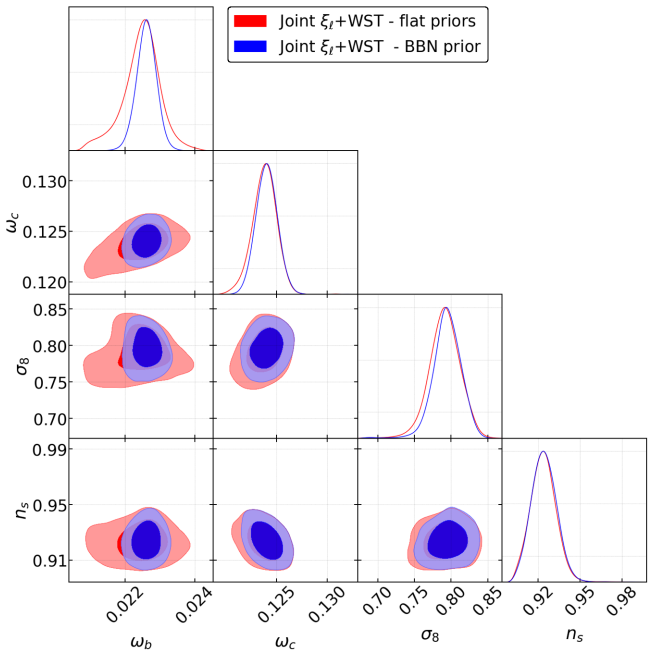

FIG. 16. Marginalized constraints on the ΛCDM cosmological parameters obtained from the joint WST + correlation function analysis using a flat prior on the value of ω_b (red), as opposed to the main analysis that used a Gaussian prior (24), in blue (and originally presented in Fig. 8).

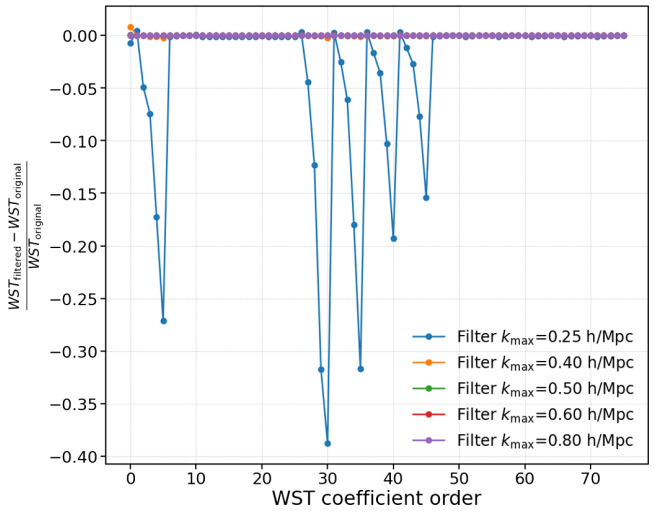

FIG. 17. Fractional changes to the WST data vector when a sharp top-hat filter with various $k_{\rm max}$ cutoff values is applied to the galaxy field before the evaluation in Eq. (6), with respect to the original evaluation of our main analysis. This example evaluation corresponds to the NGC part of the BOSS dataset.

potential undesired contributions from higher frequencies (smaller scales). In Fig. 17, we plot the fractional change to the WST data vector obtained from the BOSS NGC data when this filtering is applied for various cutoffs, compared to the original prediction using just the Gaussian-like smoothing from Eq. (4), with $\sigma = 0.80$ and $N_{\rm grid} = 270$. Our 2-point correlation function benchmark analysis includes scales down to 8 Mpc/h, corresponding to a $k_{\rm max} = 0.8$ h/Mpc in the Fourier space. As we see in Fig. 17, imposing this sharp cut-off leads to no measurable changes in the WST data vector compared to the original one, indicating no sensitivity to $k > 0.8$ h/Mpc. Imposing progressively stricter cutoff values leads to growing differences in the data vector, as we remove scales that our wavelets were originally sensitive to. If we restrict our focus on wavemodes $k \leq 0.25$ h/Mpc, the changes in the data vector are the most pronounced, as expected, since that would discard the majority of the nonlinear information contained in the galaxy field. Different combinations of the Gaussian width and/or grid size lead to a different spatial support, which can similarly be further contained with the sharp k-space filter. We also confirmed that the behavior in Fig. 17 holds not just for the BOSS data, but also for our simulation-based model predictions across the prior space. These findings confirm that our specific choices of Gaussian width, grid size and harmonic order for the wavelet analysis were conservative enough and did not access scales smaller than the ones of the correlation function benchmark analysis.

We also note that wavelets which are explicitly designed to have a finite support in Fourier space, such as e.g. the ones used in [83,84], are a natural next improvement to the above approach, that we are actively working on implementing in advance of the application to the next generation of spectroscopic data.

[1] M. Levi *et al.* (DESI Collaboration), The DESI experiment, a white paper for Snowmass 2013, arXiv:1308.0847.
[2] A. Aghamousa, J. Aguilar, S. Ahlen, S. Alam, L. E. Allen, C. Allende Prieto, J. Annis, S. Bailey, C. Balland, O. Ballester, C. Baltay, L. Beaufore, C. Bebek, T. C. Beers, E. F. Bell *et al.* (DESI Collaboration), The DESI experiment part I: Science, targeting, and survey design, arXiv:1611.00036.
[3] P. A. Abell *et al.* (LSST Science and LSST Project Collaborations), LSST science book, version 2.0, arXiv:0912.0201.
[4] A. Abate *et al.* (LSST Dark Energy Science Collaboration), Large synoptic survey telescope: Dark Energy Science Collaboration, arXiv:1211.0310.
[5] R. Laureijs *et al.* (EUCLID Collaboration), Euclid definition study report, arXiv:1110.3193.
[6] 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.
[7] E. J. Copeland, M. Sami, and S. Tsujikawa, Dynamics of dark energy, Int. J. Mod. Phys. D **15**, 1753 (2006).
[8] 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.
[9] M. Ishak, Testing general relativity in cosmology, Living Rev. Relativity **22**, 1 (2019).
[10] P. G. Ferreira, Cosmological tests of gravity, Annu. Rev. Astron. Astrophys. **57**, 335 (2019).
[11] S. Alam *et al.*, Testing the theory of gravity with DESI: Estimators, predictions and simulation requirements, J. Cosmol. Astropart. Phys. 11 (2021) 050.
[12] J. Lesgourgues and S. Pastor, Massive neutrinos and cosmology, Phys. Rep. **429**, 307 (2006).
[13] C. Dvorkin *et al.*, Neutrino mass from cosmology: Probing physics beyond the standard model, arXiv:1903.03689.
[14] 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.
[15] N. DePorzio, W. L. Xu, J. B. Muñoz, and C. Dvorkin, Finding eV-scale light relics with cosmological observables, Phys. Rev. D **103**, 023504 (2021).
[16] W. L. Xu, J. B. Muñoz, and C. Dvorkin, Cosmological constraints on light but massive relics, Phys. Rev. D **105**, 095029 (2022).
[17] J. Carron, Information escaping the correlation hierarchy of the convergence field in the study of cosmological parameters, Phys. Rev. Lett. **108**, 071301 (2012).
[18] 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).
[19] 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 (2016).
[20] F. Bernardeau, S. Colombi, E. Gaztañaga, and R. Scoccimarro, Large-scale structure of the universe and cosmological perturbation theory, Phys. Rep. **367**, 1 (2002).
[21] 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.
[22] 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.
[23] O. H. E. Philcox and M. M. Ivanov, Boss DR12 full-shape cosmology:ΛCDM constraints from the large-scale galaxy power spectrum and bispectrum monopole, Phys. Rev. D **105**, 043517 (2022).
[24] S.-F. Chen, H. Lee, and C. Dvorkin, Precise and accurate cosmology with CMB × LSS power spectra and bispectra, J. Cosmol. Astropart. Phys. 05 (2021) 030.
[25] 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.
[26] C. Hahn, M. Eickenberg, S. Ho, J. Hou, P. Lemos, E. Massara, C. Modi, A. Moradinezhad Dizgah, L. Parker, and B. Régaldo-Saint Blancard, SIMBIG: The first cosmological constraints from the non-linear galaxy bispectrum, arXiv:2310.15243.
[27] V. Ajani, M. Baldi, A. Barthelemy, A. Boyle, P. Burger, V. F. Cardone, S. Cheng, S. Codis, C. Giocoli, J. Harnois-Déraps, S. Heydenreich, V. Kansal, M. Kilbinger, L. Linke, C. Llinares *et al.* (Euclid Collaboration), Euclid preparation XXIX: Forecasts for 10 different higher-order weak lensing statistics, Astron. Astrophys. **675**, A120 (2023).
[28] M. Schmittfull, T. Baldauf, and U. Seljak, Near optimal bispectrum estimators for large-scale structure, Phys. Rev. D **91**, 043530 (2015).
[29] 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.
[30] P. Chakraborty, S.-F. Chen, and C. Dvorkin, Skewing the CMB × LSS: A fast method for bispectrum analysis, J. Cosmol. Astropart. Phys. 07 (2022) 038.
[31] J. Hou, A. Moradinezhad Dizgah, C. Hahn, and E. Massara, Cosmological information in skew spectra of biased tracers in redshift space, J. Cosmol. Astropart. Phys. 03 (2023) 045.
[32] S.-F. Chen, P. Chakraborty, and C. Dvorkin, Analysis of BOSS galaxy data with weighted skew-spectra, arXiv:2401.13036.
[33] J. Hou, A. Moradinezhad Dizgah, C. Hahn, M. Eickenberg, S. Ho, P. Lemos, E. Massara, C. Modi, L. Parker, and B. Régaldo-Saint Blancard, SIMBIG: Cosmological constraints from the redshift-space galaxy skew spectra, arXiv:2401.15074.
[34] A. Pisani *et al.*, Cosmic voids: A novel probe to shed light on our Universe, arXiv:1903.05161.

[35] E. Massara, F. Villaescusa-Navarro, M. Viel, and P. Sutter, Voids in massive neutrino cosmologies, J. Cosmol. Astropart. Phys. 11 (2015) 018.
[36] 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).
[37] Y.-C. Cai, N. Padilla, and B. Li, Testing gravity using cosmic voids, Mon. Not. R. Astron. Soc. **451**, 1036 (2015).
[38] 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.
[39] 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).
[40] T. Bonnaire, N. Aghanim, J. Kuruvilla, and A. Decelle, Cosmology with cosmic web environments I. Real-space power spectra, Astron. Astrophys. **661**, A146 (2022).
[41] S. Radinović, S. Nadathur, H. A. Winther, W. J. Percival, A. Woodfinden, E. Massara, E. Paillas, S. Contarini, N. Hamaus, A. Kovacs, A. Pisani, G. Verza, M. Aubert, A. Amara, N. Auricchio, M. Baldi *et al.*, Euclid: Cosmology forecasts from the void-galaxy cross-correlation function with reconstruction, Astron. Astrophys. **677**, A78 (2023).
[42] 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).
[43] 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).
[44] M. White, A marked correlation function for constraining modified gravity models, J. Cosmol. Astropart. Phys. 11 (2016) 057.
[45] G. Valogiannis and R. Bean, Beyond δ: Tailoring marked statistics to reveal modified gravity, Phys. Rev. D **97**, 023535 (2018).
[46] 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).
[47] E. Massara, F. Villaescusa-Navarro, C. Hahn, M. M. Abidi, M. Eickenberg, S. Ho, P. Lemos, A. Moradinezhad Dizgah, and B. Régaldo-Saint Blancard, Cosmological information in the marked power spectrum of the galaxy field, Astrophys. J. **951**, 70 (2023).
[48] E. Paillas, Y.-C. Cai, N. Padilla, and A. G. Sánchez, Redshift-space distortions with split densities, Mon. Not. R. Astron. Soc. **505**, 5731 (2021).
[49] E. Paillas, C. Cuesta-Lazaro, P. Zarrouk, Y.-C. Cai, W. J. Percival, S. Nadathur, M. Pinon, A. de Mattia, and F. Beutler, Constraining $\nu\Lambda$CDM with density-split clustering, Mon. Not. R. Astron. Soc. **522**, 606 (2023).
[50] A. E. Bayer, F. Villaescusa-Navarro, E. Massara, J. Liu, D. N. Spergel, L. Verde, B. Wandelt, M. Viel, and S. Ho, Detecting neutrino mass by combining matter clustering, halos, and voids, Astrophys. J. **919**, 24 (2021).
[51] E. Paillas, C. Cuesta-Lazaro, W. J. Percival, S. Nadathur, Y.-C. Cai, S. Yuan, F. Beutler, A. de Mattia, D. Eisenstein, D. Forero-Sanchez, N. Padilla, M. Pinon, V. Ruhlmann-Kleider, A. G. Sánchez, G. Valogiannis, and P. Zarrouk, Cosmological constraints from density-split clustering in the BOSS CMASS galaxy sample, arXiv:2309.16541.
[52] C. Cuesta-Lazaro, E. Paillas, S. Yuan, Y.-C. Cai, S. Nadathur, W. J. Percival, F. Beutler, A. de Mattia, D. Eisenstein, D. Forero-Sanchez, N. Padilla, M. Pinon, V. Ruhlmann-Kleider, A. G. Sánchez, G. Valogiannis, and P. Zarrouk, SUNBIRD: A simulation-based model for full-shape density-split clustering, arXiv:2309.16539.
[53] A. Banerjee and T. Abel, Nearest neighbour distributions: New statistical measures for cosmological clustering, Mon. Not. R. Astron. Soc. **500**, 5479 (2020).
[54] A. Banerjee and T. Abel, Cosmological cross-correlations and nearest neighbour distributions, Mon. Not. R. Astron. Soc. **504**, 2911 (2021).
[55] 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).
[56] G. Pratten and D. Munshi, Non-Gaussianity in large-scale structure and Minkowski functionals, Mon. Not. R. Astron. Soc. **423**, 3209 (2012).
[57] 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).
[58] W. Fang, B. Li, and G.-B. Zhao, New probe of departures from general relativity using Minkowski functionals, Phys. Rev. Lett. **118**, 181301 (2017).
[59] W. Liu, A. Jiang, and W. Fang, Probing massive neutrinos with the Minkowski functionals of the galaxy distribution, J. Cosmol. Astropart. Phys. 09 (2023) 037.
[60] K. Naidoo, E. Massara, and O. Lahav, Cosmology and neutrino mass with the minimum spanning tree, Mon. Not. R. Astron. Soc. **513**, 3596 (2022).
[61] 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).
[62] D. Jamieson and M. Loverde, Position-dependent matter density probability distribution function, Phys. Rev. D **102**, 123546 (2020).
[63] J. Bruna and S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1872 (2013).
[64] 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).
[65] 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).
[66] 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, Astrophys. J. **954**, 11 (2023).
[67] C. Dvorkin *et al.*, Machine learning and cosmology, in *2022 Snowmass Summer Study* (2022), arXiv:2203.08056.
[68] C. Hahn *et al.*, SIMBIG: The first cosmological constraints from non-Gaussian and non-linear galaxy clustering, arXiv:2310.15246.
[69] S. Mallat, Group invariant scattering, Commun. Pure Appl. Math. **65**, 1331 (2012).
[70] 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)* (2013), https://ieeexplore.ieee.org/document/6619007.
[71] J. Bruna, S. Mallat, E. Bacry, and J.-F. Muzy, Intermittent process analysis with scattering moments, Ann. Stat. **43**, 323 (2015).
[72] J. Andén and S. Mallat, Deep scattering spectrum, IEEE Trans. Signal Process. **62**, 4114 (2014).
[73] S. Cheng and B. Menard, How to quantify fields or textures? A guide to the scattering transform, arXiv:2112.01288.
[74] 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).
[75] 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).
[76] 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).
[77] 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).
[78] S. Cheng and B. Ménard, Weak lensing scattering transform: Dark energy and neutrino mass sensitivity, Mon. Not. R. Astron. Soc. **507**, 1012 (2021).
[79] G. Valogiannis and C. Dvorkin, Towards an optimal estimation of cosmological parameters with the wavelet scattering transform, Phys. Rev. D **105**, 103534 (2022).
[80] G. Valogiannis and C. Dvorkin, Going beyond the galaxy power spectrum: An analysis of boss data with wavelet scattering transforms, Phys. Rev. D **106**, 103509 (2022).
[81] 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).
[82] B. Greig, Y.-S. Ting, and A. A. Kaurov, Exploring the cosmic 21-cm signal from the epoch of reionization using the wavelet scattering transform, Mon. Not. R. Astron. Soc. **513**, 1719 (2022).
[83] 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.
[84] B. Regaldo-Saint Blancard, C. Hahn, S. Ho, J. Hou, P. Lemos, E. Massara, C. Modi, A. Moradinezhad Dizgah, L. Parker, Y. Yao, and M. Eickenberg, SIMBIG: Galaxy clustering analysis with the wavelet scattering transform, arXiv:2310.15250.
[85] D. T. Chung, Exploration of 3D wavelet scattering transform coefficients for line-intensity mapping measurements, Mon. Not. R. Astron. Soc. **517**, 1625 (2022).
[86] 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.
[87] 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).
[88] 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 quijote simulations, Astrophys. J. Suppl. Ser. **250**, 2 (2020).
[89] 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).
[90] 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 (2012).
[91] 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).
[92] 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.
[93] 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).
[94] 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).
[95] 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 (2016).
[96] 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.
[97] 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 (2022).
[98] 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).
[99] 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).
[100] M. Levi, C. Bebek, T. Beers, R. Blum, R. Cahn, D. Eisenstein, B. Flaugher, K. Honscheid, R. Kron, O. Lahav, P. McDonald, N. Roe, D. Schlegel (The DESI Collaboration), The DESI experiment, a whitepaper for Snowmass 2013, arXiv:1308.0847.
[101] 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 (2022).
[102] 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).
[103] S. Yuan, L. H. Garrison, D. J. Eisenstein, and R. H. Wechsler, Stringent σ_8 constraints from small-scale galaxy clustering using a hybrid MCMC + emulator framework, Mon. Not. R. Astron. Soc. **515**, 871 (2022).
[104] E. Calabrese, R. A. Hložek, J. R. Bond, M. J. Devlin, J. Dunkley, M. Halpern, A. D. Hincks, K. D. Irwin, A. Kosowsky, K. Moodley, L. B. Newburgh, M. D. Niemack, L. A. Page, B. D. Sherwin, J. L. Sievers, D. N. Spergel *et al.*, Cosmological parameters from pre-Planck CMB measurements: A 2017 update, Phys. Rev. D **95**, 063525 (2017).
[105] Z. Zheng, A. A. Berlind, D. H. Weinberg, A. J. Benson, C. M. Baugh, S. Cole, R. Davé, C. S. Frenk, N. Katz, and C. G. Lacey, Theoretical models of the halo occupation distribution: Separating central and satellite galaxies, Astrophys. J. **633**, 791 (2005).
[106] 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).
[107] H. Guo, Z. Zheng, I. Zehavi, K. Dawson, R. A. Skibba, J. L. Tinker, D. H. Weinberg, M. White, and D. P. Schneider, Velocity bias from the small-scale clustering of SDSS-III BOSS galaxies, Mon. Not. R. Astron. Soc. **446**, 578 (2015).
[108] S. Yuan, L. H. Garrison, B. Hadzhiyska, S. Bose, and D. J. Eisenstein, AbacusHOD: A highly efficient extended multi-tracer HOD framework and its application to BOSS and eBOSS data, Mon. Not. R. Astron. Soc. **510**, 3301 (2021).
[109] 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).
[110] J.-N. Ye, H. Guo, Z. Zheng, and I. Zehavi, Properties and origin of galaxy velocity bias in the illustris simulation, Astrophys. J. **841**, 45 (2017).
[111] 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).
[112] https://github.com/abacusorg/abacusutils.
[113] https://abacusutils.readthedocs.io/en/latest/hod.html.
[114] H. A. Feldman, N. Kaiser, and J. A. Peacock, Power-spectrum analysis of three-dimensional redshift surveys, Astrophys. J. **426**, 23 (1994).
[115] 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).
[116] 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 (2016).
[117] F. Beutler and P. McDonald, Unified galaxy power spectrum measurements from 6dfgs, BOSS, and eBOSS, J. Cosmol. Astropart. Phys. 11 (2021) 031.
[118] F. G. Mohammad, F. Villaescusa-Navarro, S. Genel, D. Angles-Alcazar, and M. Vogelsberger, Inpainting hydrodynamical maps with deep learning, Astrophys. J. **941**, 132 (2022).
[119] 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).
[120] C. Alcock and B. Paczynski, An evolution free test for non-zero cosmological constant, Nature (London) **281**, 358 (1979).
[121] M. M. Ivanov, M. Simonović, and M. Zaldarriaga, Cosmological parameters from the BOSS galaxy power spectrum, J. Cosmol. Astropart. Phys. 05 (2020) 042.
[122] 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.
[123] 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 h0, J. Cosmol. Astropart. Phys. 05 (2020) 032.
[124] 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.
[125] 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.
[126] R. Hockney and J. Eastwood, *Computer Simulation Using Particles*, Advanced Book Program (McGraw-Hill International Book Company, New York, 1981).
[127] M. Sinha and L. H. Garrison, Corrfunc—A suite of blazing fast correlation functions on the CPU, Mon. Not. R. Astron. Soc. **491**, 3022 (2020).
[128] S. D. Landy and A. S. Szalay, Bias and variance of angular correlation functions, Astrophys. J. **412**, 64 (1993).
[129] K. Heitmann, D. Higdon, M. White, S. Habib, B. J. Williams, E. Lawrence, and C. Wagner, The Coyote Universe. II. Cosmological models and precision emulation of the nonlinear matter power spectrum, Astrophys. J. **705**, 156 (2009).
[130] E. Lawrence, K. Heitmann, M. White, D. Higdon, C. Wagner, S. Habib, and B. Williams, The Coyote Universe. III. Simulation suite and precision emulator for the nonlinear matter power spectrum, Astrophys. J. **713**, 1322 (2010).
[131] K. Heitmann, E. Lawrence, J. Kwan, S. Habib, and D. Higdon, The Coyote Universe extended: Precision emulation of the matter power spectrum, Astrophys. J. **780**, 111 (2014).
[132] N. Ramachandra, G. Valogiannis, M. Ishak, K. Heitmann (LSST Dark Energy Science Collaboration), Matter power spectrum emulator for f(R) modified gravity cosmologies, Phys. Rev. D **103**, 123525 (2021).
[133] Z. Zhai, J. L. Tinker, M. R. Becker, J. DeRose, Y.-Y. Mao, T. McClintock, S. McLaughlin, E. Rozo, and R. H. Wechsler, The Aemulus project. III. Emulation of the galaxy correlation function, Astrophys. J. **874**, 95 (2019).
[134] 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.
[135] Y. Kobayashi, T. Nishimichi, M. Takada, and H. Miyatake, Full-shape cosmology analysis of the SDSS-III BOSS galaxy power spectrum using an emulator-based halo model: A 5% determination of σ_8, Phys. Rev. D **105**, 083517 (2022).
[136] S. Yuan, B. Hadzhiyska, and T. Abel, Full forward model of galaxy clustering statistics with AbacusSummit light cones, Mon. Not. R. Astron. Soc. **520**, 6283 (2023).
[137] 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).
[138] 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 catalogues for the BOSS final data release, Mon. Not. R. Astron. Soc. **456**, 4156 (2016).
[139] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, EMCEE: The MCMC hammer, Publ. Astron. Soc. Pac. **125**, 306 (2013).
[140] 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).
[141] 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).
[142] 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).
[143] A. Klypin and J. Holtzman, Particle mesh code for cosmological simulations, arXiv:astro-ph/9712217.
[144] A. V. Kravtsov, A. A. Berlind, R. H. Wechsler, A. A. Klypin, S. Gottlober, B. Allgood, and J. R. Primack, The dark side of the halo occupation distribution, Astrophys. J. **609**, 35 (2004).
[145] D. Kodwani, D. Alonso, and P. Ferreira, The effect on cosmological parameter estimation of a parameter dependent covariance matrix, Open J. Astrophys. **2** (2019), 10.21105/astro.1811.11584.
[146] J. Carron, On the assumption of Gaussianity for cosmological two-point statistics and parameter dependent covariance matrices, Astron. Astrophys. **551**, A88 (2013).
[147] O. H. E. Philcox and J. Ereza, Could sample variance be responsible for the parity-violating signal seen in the BOSS galaxy survey?, arXiv:2401.09523.
[148] T. Ishiyama, F. Prada, A. A. Klypin, M. Sinha, R. B. Metcalf, E. Jullo, B. Altieri, S. A. Cora, D. Croton, S. de la Torre, D. E. Millán-Calero, T. Oogi, J. Ruedas, and C. A. Vega-Martínez, The Uchuu simulations: Data Release 1 and dark matter halo concentrations, Mon. Not. R. Astron. Soc. **506**, 4210 (2021).
[149] C. A. Dong-Páez, A. Smith, A. O. Szewciw, J. Ereza, M. H. Abdullah, C. Hernández-Aguayo, S. Trusov, F. Prada, A. Klypin, T. Ishiyama, A. Berlind, P. Zarrouk, J. López Cacheiro, and J. Ruedas, The Uchuu-SDSS galaxy lightcones: A clustering, RSD and BAO study, Mon. Not. R. Astron. Soc. **528**, 7236 (2024).

[150] T. Oogi, T. Ishiyama, F. Prada, M. Sinha, D. Croton, S. A. Cora, E. Jullo, A. A. Klypin, M. Nagashima, J. López Cacheiro, J. Ruedas, M. A. R. Kobayashi, and R. Makiya, Uchuu-ν^2GC galaxies and AGN: Cosmic variance forecasts of high-redshift AGN for JWST, Euclid, and LSST, Mon. Not. R. Astron. Soc. **525**, 3879 (2023).

[151] H. Aung, D. Nagai, A. Klypin, P. Behroozi, M. H. Abdullah, T. Ishiyama, F. Prada, E. Pérez, J. López Cacheiro, and J. Ruedas, The Uchuu-universe machine data set: Galaxies in and around clusters, Mon. Not. R. Astron. Soc. **519**, 1648 (2023).

[152] F. Prada, P. Behroozi, T. Ishiyama, A. Klypin, and E. Pérez, Confirmation of the standard cosmological model from red massive galaxies ∼600 Myr after the big bang, arXiv:2304.11911.

[153] T. Ishiyama, T. Fukushige, and J. Makino, GreeM: Massively parallel TreePM code for large cosmological N -body simulations, Publ. Astron. Soc. Jpn. **61**, 1319 (2009).

[154] P. S. Behroozi, C. Conroy, and R. H. Wechsler, A comprehensive analysis of uncertainties affecting the stellar mass–halo mass relation for $0 < z < 4$, Astrophys. J. **717**, 379 (2010).

[155] P. Behroozi, R. H. Wechsler, A. P. Hearin, and C. Conroy, UniverseMachine: The correlation between galaxy growth and dark matter halo assembly from $z = 0$-10, Mon. Not. R. Astron. Soc. **488**, 3143 (2019).

[156] S. Alam *et al.* (BOSS Collaboration), The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: Cosmological analysis of the DR12 galaxy sample, Mon. Not. R. Astron. Soc. **470**, 2617 (2017).

[157] 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, Astrophys. J. **948**, 99 (2023).

[158] M. J. Chapman *et al.*, The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Measurement of the growth rate of structure from the small-scale clustering of the luminous red galaxy sample, Mon. Not. R. Astron. Soc. **516**, 617 (2022).

[159] J. E. Bautista *et al.*, The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Measurement of the BAO and growth rate of structure of the luminous red galaxy sample from the anisotropic correlation function between redshifts 0.6 and 1, Mon. Not. R. Astron. Soc. **500**, 736 (2020).

[160] A. de Mattia *et al.*, The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Measurement of the BAO and growth rate of structure of the emission line galaxy sample from the anisotropic power spectrum between redshift 0.6 and 1.1, Mon. Not. R. Astron. Soc. **501**, 5616 (2021).

[161] B. Yu, U. Seljak, Y. Li, and S. Singh, Rsd measurements from boss galaxy power spectrum using the halo perturbation theory model, J. Cosmol. Astropart. Phys. 04 (2023) 057.

[162] F. Beutler, C. Blake, M. Colless, D. H. Jones, L. Staveley-Smith, G. B. Poole, L. Campbell, Q. Parker, W. Saunders, and F. Watson, The 6dF galaxy survey: $Z \approx 0$ measurements of the growth rate and σ_8Mon. Not. R. Astron. Soc. **423**, 3430 (2012).

[163] S. Yuan, T. Abel, and R. H. Wechsler, Robust cosmological inference from non-linear scales with k-th nearest neighbor statistics, Mon. Not. R. Astron. Soc. **527**, 1993 (2023).

[164] K. S. Dawson *et al.*, The SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Overview and early data, Astron. J. **151**, 44 (2016).

[165] 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.

[166] B. Hadzhiyska *et al.*, Synthetic light cone catalogues of modern redshift and weak lensing surveys with Abacus-Summit, arXiv:2305.11935.

[167] A. K. Saydjari and D. P. Finkbeiner, Equivariant wavelets: Fast rotation and translation invariant wavelet scattering transforms, arXiv:2104.11244.

[168] O. Friedrich, F. Andrade-Oliveira, H. Camacho, O. Alves, R. Rosenfeld, J. Sanchez, X. Fang, T. F. Eifler, E. Krause, C. Chang, Y. Omori, A. Amon, E. Baxter, J. Elvin-Poole, D. Huterer, A. Porredon *et al.*, Dark Energy Survey year 3 results: Covariance modelling and its impact on parameter estimation and quality of fit, Mon. Not. R. Astron. Soc. **508**, 3125 (2021).

[169] 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, Astrophys. J. **946**, 107 (2023).