Full forward model of galaxy clustering statistics with simulation lightcones

Sihan Yuan,1,2 Boryana Hadzhiyska3,4, and Tom Abel1,2

1Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, CA 94305, USA
2Department of Particle Physics and Astrophysics, SLAC National Accelerator Laboratory, Stanford, CA 94305, USA
3Department of Physics, University of California, Berkeley, CA 94720
4Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Novel summary statistics beyond the standard 2-point correlation function (2PCF) are necessary to capture the full astrophysical and cosmological information from the small-scale ($r < 30h^{-1}$Mpc) galaxy clustering. However, the analysis of beyond-2PCF statistics on small scales is challenging because we lack the appropriate treatment of observational systematics for arbitrary summary statistics of the galaxy field. In this paper, we develop a full forward modeling pipeline for any summary statistics using high-fidelity simulation lightcones that accounts for all observational systematics and is appropriate for a wide range of summary statistics. We apply our forward model approach to a fully realistic mock galaxy catalog and demonstrate that we can recover unbiased constraints on the underlying galaxy–halo connection model using two separate summary statistics: the standard 2PCF and the novel $k$-th nearest neighbor ($k$NN) statistics, which are sensitive to correlation functions of all orders. We expect that applying this forward model approach to current and upcoming surveys while leveraging a multitude of summary statistics will become a powerful technique in maximally extracting information from the non-linear scales.

Key words: cosmology: large-scale structure of Universe – galaxies: haloes – methods: statistical – methods: numerical

1 INTRODUCTION

The spatial distribution of galaxies presents one of the most powerful probes of the fundamental properties of the universe. Over the last few decades, galaxy clustering has emerged as an essential tool in constraining cosmology and galaxy evolution, especially with the advent of wide-field spectroscopic surveys such as the SDSS-III Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013), the SDSS-IV extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2016), and the ongoing the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2016).

The standard approach to extracting cosmological information from galaxy clustering is through standard rulers and compressed statistics, most notably the baryon acoustic oscillation peak (BAO; Eisenstein et al. 2005) and the Alcock–Paczynski (AP) effect (Alcock & Paczynski 1979) measured from the 2-point correlation function (2PCF). While the standard techniques are robust to most observational systematics (Ross et al. 2012, 2015), they are also very limited in the amount of information they can extract from the data. One key limitation is that such techniques are limited to large scales, as the theory templates rely on perturbation theories, which are only reliable beyond approximately $30–50h^{-1}$Mpc (Carlson et al. 2009, 2013). However, modern cosmological surveys are designed in a way that their galaxy clustering measurements are most accurate at scales of a few megaparsecs, far below the limits of perturbative models. The small scales are also important because they are highly sensitive to non-linear growth, and thus turn out to be significantly more constraining on cosmic growth history. The other key limit of the standard approaches is that they only rely on compressions of the 2PCF, which is itself a compression of the full density field. Thus, to fully take advantage of the information content of modern cosmological surveys, we need to develop accurate and unbiased models for statistics beyond the 2PCF, and on scales extending deep into the non-linear regime.

Modeling structure on small scales is challenging. Perturbative models fail because small-scale structure is dominated by high-order contributions of both the density and velocity fields, plus non-perturbative effects arising from the dynamics beyond shell crossing, i.e., formation and evolution of galaxies (or dark matter halos) and baryonic feedback. As an alternative, a new class of models have arisen in recent years leveraging large N-body simulations instead of analytical approaches. Simulations can precisely capture the non-linear evolution of dark matter density and velocity fields, given sufficient computational resources. However, N-body simulations only simulate the gravitational growth of the total matter field and need to be paired with a robust galaxy–dark matter connection model that populates galaxies on top of the simulated matter density field. A series of recent studies have attempted to obtain cosmological constraints using simulation-based models (e.g. Zhai et al. 2019; Lange et al. 2022; Chapman et al. 2021; Kobayashi et al. 2022; Yuan et al. 2022a).

At the same time, there have been numerous studies demonstrating the significant information gain when incorporating beyond-2PCF statistics. Perhaps the most commonly discussed alternative statistics
are the 3-point correlation function (3PCF) and its Fourier counterpart, the bispectrum. Several studies have successfully applied the 3PCF/bispectrum to data and obtained cosmological constraints (Slepian et al. 2017b,a; Gil-Marín et al. 2017; D’Amico et al. 2022). However, these analyses are still limited to linear scales. Yuan et al. (2018) demonstrated the diverse information content of the squeezed 3PCF on small scales but have yet to apply it to data.

Other extensions to the 2PCF include the marked correlation function, which weights the 2PCF with a secondary tag, such as the environment, to highlight different aspects of clustering, thus complementing the vanilla 2PCF. A series of studies have found the marked 2PCF to potentially powerful in constraining the cosmological parameter $\sigma_8$ and modifications to general relativity Sheth & van de Weygaert (2004); White & Padmanabhan (2009); White (2016). On the small-scale front, Storey-Fisher et al. (2022) forecasts the cosmological constraining power of marked statistics and void statistics in a simulation-based mock analysis and found that these statistics can bring up to 30% improvements to the constraints on parameters $\sigma_8$ and $\Omega_m$.

Besides correlation-based statistics, density-based statistics such as the $k$-th nearest neighbor statistics ($kNN$) and wavelet scattering transforms (WST) have also generated considerable interest in recent years. Banerjee & Abel (2021a,b) found that the $kNN$s can break degeneracies in the 2-point clustering and potentially improve the cosmology constraints by a factor of a few. Valogiannis & Dvorkin (2022b,a) applied WST to BOSS galaxies in a preliminary analysis of the large scales and found substantial improvement in cosmological parameter constraints.

None of these studies push their beyond-2PCF analysis to small scales on data because they lack not only a robust theory template for such scales, but also a proper treatment of the effects of data systematics. Such systematics include: (1) redshift-dependent completeness, where the galaxy sample’s number density varies as a function of redshift due to survey selection cuts; (2) survey masks and geometry, where galaxies in certain regions of the survey fail to be observed due to survey windows and various foreground obstructions; and last but not least (3) fiber collision, where galaxies that are too close to each other in projection do not always get redshift measurements. Ross et al. (2012) details these systematic effects in SDSS/BOSS, but these effects are generic to all fiber-fed spectroscopic surveys, including all ground-based experiments such as eBOSS, DESI, and PFS.

The effects of fiber collision are particularly troublesome as they are density dependent and propagate to large scales. Traditionally, one can correct for the effects of missing redshifts through a weighting scheme that essentially assigns the weight of the missing galaxy to its nearest neighbor, as was done for BOSS (Anderson et al. 2014; Reid et al. 2016). However, while this correction works well on large scales, it introduces non-negligible spurious signal on small scales (Guo et al. 2012; Li et al. 2006). More sophisticated correction schemes have since been developed that can produce unbiased corrections for the 2PCF down to small scales (e.g. Mohammad et al. 2020; Smith et al. 2019; Bianchi et al. 2018), but these techniques are based on recovering pair counts instead of recovering individual redshifts, thus they are only suited for the 2PCF. Because of these limitations, no study has actually been to able to extract the rich information in the precise small-scale clustering measurements with beyond-2PCF statistics and a proper accounting of systematics.

In this paper, we formulate a full forward model approach that promises to finally enable analyzing small-scale measurements with any galaxy clustering statistics. We use state-of-the-art simulations to produce precise models of small-scale structure. We cast these simulations on lightcones to enable full accounting of the aforementioned survey systematics in a forward model framework. We demonstrate the efficacy of this approach by performing model parameter recovery using two sets of clustering statistics on a mock galaxy catalog that has all the systematics built in. We show that such a forward model approach is computationally feasible with currently available simulation products and typical computational resources. We also compare several different summary statistics in their constraining power. We encourage further development of similar ideas and methods to bring forth a zoo of robust analyses of small-scale data with new clustering statistics.

Specifically, the paper is structured as follows: In Section 2, we describe the tools and steps necessary in such a forward model. In Section 3, we have a dedicated discussion of our proposed treatment of fiber collision. In Section 4, we perform mock parameter recovery with with two different clustering statistics, the 2PCF and the $k$-th nearest neighbor cumulative distribution function ($kNN-CDF$). In Section 5, we expand on this analysis and look beyond to a full cosmology analysis on data. Finally, we conclude in Section 6.

2 FORWARD MODEL WITH SIMULATION LIGHTCONES

In this section, we describe the forward modeling approach with simulation lightcones in detail. Figure 1 provides a cartoon overview...
of the full forward model approach and serves as a guide through this section. We start with simulation lightcones, which we populate with mock galaxies. Then we apply layers of observational systematics including redshift selection \((n(z))\) and survey windows/masks. Finally we compute the desired summaries statistics, which can then be compared with data for likelihood analyses. We describe each of these steps in detail in the following subsections.

2.1 AbacusSummit lightcones

The AbacusSummit simulation suite (Maksimova et al. 2021) is a set of large, high-accuracy cosmological N-body simulations using the Abacus N-body code (Garrison et al. 2019, 2021), designed to meet and exceed the Cosmological Simulation Requirements of the Dark Energy Spectroscopic Instrument (DESI) survey (Levi et al. 2013). AbacusSummit consists of over 150 simulations, containing approximately 60 trillion particles at 97 different cosmologies. For this analysis, we use exclusively the “base” configuration boxes within the simulation suite, each of which contains 6912\(^3\) particles within a \((2h^{-1}\text{Gpc})^3\) volume, corresponding to a particle mass of \(2.1 \times 10^9h^{-1}\text{M}_\odot\). The AbacusSummit suite also uses a specialized spherical-overdensity based halo finder known as COMPASO (Hadhziyska et al. 2022a).

In addition to periodic boxes, the simulation suite also provides a set of simulation lightcones at fiducial cosmology (Hadhziyska et al. 2022b). The basic algorithm associates the halos from a set of coarsely-spaced snapshots with their positions at the time of lightcone crossing by matching halo particles to on-the-fly light cone particles. The resulting halo catalogs provide accurate interpolated position for all available “cleaned” halos in the simulation and are particularly reliable for halos with masses above \(M_{\text{halo}} \geq 1 \times 10^{11}h^{-1}\text{M}_\odot\), which is more than sufficient for the purposes of current redshift surveys. Unlike other methods, which commonly adopt a “cookie-cutting” strategy of selecting halos at their momentary positions and thus ignore the non-negligible distance traversed by halos between redshift epochs, the lightcone catalogues of AbacusSummit provide interpolation that has been shown to be accurate to less than a percent (for a more detailed discussion, see Section 4 in Hadhziyska et al. 2022b). For this analysis, we utilize the 25 base lightcones, which are constructed from the 25 base periodic boxes, with each lightcone covering an octant of the sky \((\sim 5156 \text{deg}^2)\) up to \(z \sim 0.8\).

The CMASS sample we consider for this analysis is approximately 9000 \text{deg}^2 in area and spans redshift range \(0.45 < z < 0.6\). Thus, we can exceed the data volume with two base lightcones, though ideally we want to use even more lightcones to further reduce model sample variance.

2.2 AbacusHOD on lightcone

The first step of the forward model is to populate the simulation lightcones with galaxies. To achieve this, we use a Halo Occupation Distribution (HOD; e.g. Zheng et al. 2005, 2007) approach, which probabilistically populate dark matter halos with galaxies according to a set of halo properties. For a Luminous Red Galaxy (LRG) sample, the HOD is well approximated by a vanilla model given by (originally shown in Kwan et al. (2015)):

\[
\hat{n}_{\text{cen}}(z) = \frac{ic}{2} \text{erf} \left( \frac{\log_{10}(M_{\text{cut}}/M)}{\sqrt{2}\sigma} \right),
\]

\[
\hat{n}_{\text{sat}}(z) = \left[ \frac{M - kM_{\text{cut}}}{M_1} \right]^{\alpha}\hat{n}_{\text{cen}}(M),
\]

where the five vanilla parameters characterizing the model are \(M_{\text{cut}}, M_1, \sigma, \alpha, k\). \(M_{\text{cut}}\) characterizes the minimum halo mass that host a central galaxy. \(M_1\) characterizes the typical halo mass that hosts one satellite galaxy. \(\sigma\) describes the steepness of the transition from 0 to 1 in the number of central galaxies. \(\alpha\) is the power law index on the number of satellite galaxies. \(kM_{\text{cut}}\) gives the minimum halo mass to host a satellite galaxy. We have added a modulation term \(\hat{n}_{\text{cen}}(M)\) to the satellite occupation function to remove satellites from halos without centrals. We have also included an incompleteness parameter \(ic\), which is a downsampling factor controlling the overall number density of the mock galaxies. This parameter is relevant when trying to match the observed mean density of the galaxies in addition to clustering measurements. By definition, \(0 < ic \leq 1\).

The lightcone also enables modeling of galaxy population across redshifts, where both the underlying density field and the galaxy bias evolves. This is particularly powerful for current and future deep surveys such as DESI and PFS where simultaneous modeling across a wide redshift range is necessary to leverage the full statistical power of the data.

To this end, we enable redshift-dependent HOD models in AbacusHOD. In a first implementation, we add two additional free parameters to the model, \(\mu_{\text{cut},p}\) and \(\mu_{1,p}\), where we define \(\mu \equiv \log M\). These two parameters are the first derivative of \(\log M_{\text{cut}}\) and \(\log M_1\) against the scale factor, respectively. Thus, the modified parameters in the AbacusHOD model take the following form for a given choice of a reference redshift, \(z_{\text{pivot}}\):

\[
\log M_i(z) = \log M_i(z_{\text{pivot}}) + \mu_i \left( \frac{1}{1+z} - \frac{1}{1+z_{\text{pivot}}} \right),
\]

where \(i = \{\text{cut}, 1\}\). Currently, we opt to make only \(M_{\text{cut}}\) and \(M_1\) redshift-dependent for simplicity. In principle, all HOD parameters can be redshift-dependent and the AbacusHOD package can be easily extended to accommodate such complexities.

In addition to determining the number of galaxies per halo, the standard HOD model also dictates the position of velocity of the galaxies. For the central galaxy, its position and velocity are set to be the same as those the halo center, specifically the L2 subhalo center-of-mass for the COMPASO halos. For the satellite galaxies, they are randomly assigned to halo particles with uniform weights, each satellite inheriting the position and velocity of its host particle.

For this paper, we fix two parameters \(\sigma\) and \(k\) in the vanilla HOD for simplicity. \(k\) does not strongly affect clustering and only comes into effect at very small scales. \(\sigma\) does affect clustering on 2-halo scales, but it tends to be strongly degenerate with \(\log M_{\text{cut}}\). We omit \(\sigma\) in this preliminary analysis for clearer interpretation of the results. We also ignore redshift-dependence in the HOD, setting \(\mu_{\text{cut},p} = 0\) and \(\mu_{1,p} = 0\). Thus, in the following analysis, the HOD is fully parameterized by 4 parameters, \(M_{\text{cut}}, M_1, \sigma, \text{and} ic\).

In order to sample the model parameter space, each forward model step needs to be computationally efficient so as to minimize the time need to evaluate the full forward model. To this end, we adopt the highly optimized AbacusHOD implementation, which significantly speeds up the HOD calculation per HOD parameter combination (Yuan et al. 2021). The code is publicly available as a part of the abacusutils package at [https://github.com/](https://github.com/).
2.3 Survey systematics

In our forward model, we account for both redshift-dependent completeness and the survey geometry. Both of these systematic effects can significantly bias the measurements but are hard to model from a periodic simulation box. Simulation lightcones allow these effects to be modeled relatively straight-forwardly. A third critical systematic effect is fiber collision, but we reserve that discussion for Section 3.

To apply redshift-dependent completeness to our forward model, we compute the density of galaxies generated by the HOD $n_{HOD}$, and then run a filtering step where we retain each galaxy with probability $P(z) = n_{CMASS}(z)/n_{HOD}$. This step ensures the resulting number density profile mimics the observation $n_{CMASS}(z)$. This step is computationally efficient and can trivially exploit parallelization.

Due to the complex geometry of the survey boundaries, any summary statistics measurement on the full survey footprint would suffer additional systematic effects. To model such effects in a forward model, ideally one would want to generate a sufficiently large lightcone that would enclose the entire survey footprint. Then, one can trivially apply the survey selection functions on the lightcone and obtain the exact same geometry in the model as the data. However, by being density-based statistics, kNNs are also highly sensitive to systematics that result in missing objects. Such sensitivities highlight the need for a full forward model approach.

2.4 Summary statistics

Having generated the lightcone galaxy catalogs and forward-modeled the full range of systematics, one can now compute the desired summary statistics and perform likelihood analysis against data. While the forward modeling approach is fully applicable to any set of clustering statistics, we focus on the novel $k$-th nearest neighbor statistics and the more widely used 2-point correlation function in this paper. We focus on $k$NNs because they are particularly computationally efficient ($O(N \log N)$) and trivially account for high-order clustering information. However, by being density-based statistics, $k$NNs are also highly sensitive to systematics that result in missing objects.

2.4.1 $k$-th nearest neighbor statistics

In this section, we give a quick review of the $k$NN formalism and our Python implementation. For detailed derivations and illustrations, we refer the readers to Banerjee & Abel (2021a,b). To define $k$NNs, we first define $P_{k|V}$, the probability of finding exactly $k$ data points in a volume $V$, averaged over all query points inside the survey volume. We can write out $P_{k|V}$ in terms of its cumulative counterparts as

$$P_{k|V} = P_{>k-1|V} - P_{>k|V}, \forall k \geq 1,$$

where we have also reformulated the CDFs as a function of radial distance between the query point and the data point $r$. These CDFs as a function of $r$ and of order $k$ form a series of summary statistics that we later refer to as the $k$NNs. Banerjee & Abel (2021a) showed that the $k$NNs automatically includes information from all orders of correlation function without the penalty of increased functional complexity. For plotting purposes, we also define the peaked CDFs (pCDF) as

$$pCDF_k(r) = \begin{cases} kNN-CDF(r) & \text{if } kNN-CDF(r) < 0.5, \\ 1 - kNN-CDF(r) & \text{otherwise.} \end{cases}$$

Conceptually, one can think of the $k$NN-CDF as the cumulative distribution of the distances from query points to the $k$-th nearest neighbors.
The calculation of the $k$NN-CDF is straightforward. We first generate a large grid of query points. For this paper, we adopt a grid spacing of $l = 4h^{-1}\text{Mpc}$ and then only select the grid points that are within the trimmed survey volume we consider. Then we construct a $k$DTree of the 3D positions of galaxies, from which we inquire the distance to the $k$-th nearest neighbor of all the query points. Finally, we sort the distances from all the query points and construct a CDF.

It is worth noting that the choice of $l = 4h^{-1}\text{Mpc}$ effectively imposes a minimum scale for the $k$NN-CDF, as scales below grid spacing $l$ will be poorly sampled and thus carry significant noise. This point becomes important when we set up the mock test in section 4. In principle, reducing the grid spacing would allow us to access NNs on smaller scales, but at significant computational cost. We argue that this coarse spacing is sufficient for the purpose of this pilot study, but advocate for more advanced techniques for more optimal sampling of query points, such as the one proposed in Appendix A of Garrison et al. (2022).

2.4.2 2-point correlation function

We compare the $k$NN-CDF with the more commonly used 2-point correlation function in this paper. Specifically, we use projected 2PCF $w_p$:

$$w_p(r_p) = 2 \int_0^{\pi_{\text{max}}} \xi(r_p, r_\parallel) d\pi,$$

where $r_p$ and $r_\parallel$ are the transverse and line-of-sight (LoS) separations in comoving units. $\xi(r_p, r_\parallel)$ is the redshift-space 2PCF, which can be computed via the Landy & Szalay (1993) estimator:

$$\xi(r_p, r_\parallel) = \frac{DD - 2DR + RR}{RR},$$

where $DD$, $DR$, and $RR$ are the normalized numbers of data-data, data-random, and random-random pair counts in each bin of $(r_p, r_\parallel)$. For implementation, we use the highly-optimized grid-based Corrfunc code (Sinha & Garrison 2020) for fast calculations. In the case of the lightcones, which have a more complex geometry: namely, three boxes intersected by concentric shells (see Fig. 1 in Hadzhiyska et al. 2022b), we generate randoms by populating an octant of a shell of thickness determined by the lightcone crossing comoving distance for each redshift epoch, disposing of particles outside the three boxes at higher redshifts. $w_p$ is commonly used in cosmology because it marginalizes over the LOS positions of galaxies, which tend to suffer from significant redshift uncertainties, especially in the case of photometry-only data. However, the marginalization comes at the cost of losing out on the information embedded in the LoS structure of $\xi(r_p, r_\parallel)$. In Section 3, we discuss one source of such redshift uncertainty and an effective remedy for it.

2.5 Computational efficiency

So far we have discussed every step of the lightcone-based forward model except for fiber collision, which we exclude from the forward model and instead treat as a correction step on the data (Section 3). In practice, one would need to repeatedly re-evaluate the forward model to sample the posterior parameter space. To this end, it is essential to optimize the computational efficiency of the forward model, especially given the added complexity of modeling systematic effects. In this section, we report the timing and computational speed-ups we implemented for our forward modeling steps. The timing is done on a modest machine with two Intel Xeon Gold 5218 chips clocked at 2.3 GHz for a total of 32 physical cores and 256 GB DDR4-2666 RAM.

With current computational hardware, we exclude the possibility of re-running the N-body evolution to produce lightcones of sufficient volume and precision at a large number of cosmologies. In a cosmology analysis, we instead rely on interpolative methods, commonly known as emulators in cosmology, to conduct cosmology sampling given a relatively sparse set of fully-evolved cosmologies.

For a fixed cosmology, we start with generating mock galaxies on the lightcone. This step is carried out via the AbacusHOD code, which is highly optimized for parallel CPU architecture. For a single lightcone covering an octant of the sky, an CMASS-like LRG sample, each HOD evaluation takes $\sim 0.06$ seconds. This is relatively insignificant compared to later steps modeling systematics and computing summary statistics. In fact, the lightcones of AbacusSummit were designed for precisely the goal of enabling fast and accurate generation of mock catalogues on the lightcone. This was achieved by outputting them in the same format as the main halo catalogues of AbacusSummit, which allows us to use the same optimized prescription, AbacusHOD, as in the case of the cubic box.

Applying the $n(z)$ filter can be trivially parallelized and each evaluation takes $\sim 0.04$ seconds. Applying survey mask is significantly slower at $\sim 0.5$ seconds per evaluation. The performance is bottlenecked by the existing Python implementations of MANGLE$^2$ (Swanson et al. 2008; Hamilton & tegmark 2004), which is necessary to manipulate the archival BOSS survey mask files. While the application of survey masks to the galaxy catalogs can be parallelized, the run time is dominated by overhead at above $\sim 8$ threads. We note that this performance is highly specific to the BOSS survey mask files, which were developed on outdated methodologies and not optimized for performance. For upcoming datasets such as DESI, we can likely optimize the mask file formats and application algorithms for parallelization.

Finally, the summary statistics calculation takes $\sim 0.5$ seconds for each $k$NN-CDF evaluation ($\sim 0.2$ seconds for each 2PCF evaluation). The rough breakdown of the time spent per $k$NN-CDF evaluation is the following: 1) $\sim 0.1$ seconds on transforming spherical coordinates to cartesian coordinates before constructing $k$DTree; 2) $\sim 0.05$ seconds on constructing $k$DTree; 3) $\sim 0.15$ seconds on neighbor queries; 4) $\sim 0.2$ seconds on sorting and constructing CDFs. The only step that is currently fully parallelized is step 3, and it scales well with number of cores. Step 1 can be parallelized in principle for modest performance gains. Step 2 is an intrinsically serial task, and we have not come across a successful parallel implementation, but it is a relatively cheap process as it is. Step 4 can be parallelized in principle with parallel mergesort algorithms, but our tests showed insignificant performance gains when using available parallel sorting algorithms compared to NumPy Quicksort.

While the $k$NN-CDF calculation is relatively slow, it does scale well with number of radial bins (only affects step 4) and order $k$ (only affects step 3,4). Thus, $k$NN-CDF can still be computationally advantageous when compared to high-order correlation functions. We also expect to achieve significant speed-ups by building up a grid-based $k$NN calculator from scratch, adopting many of the techniques used for Corrfunc.

$^2$ https://github.com/esheldon/pymangle
3 FIBER COLLISION CORRECTION

Fiber collision refers to the effect where spectroscopic fibers are not infinitely thin so one cannot put two fibers infinitely close to each other. For example, in BOSS, the minimum angular distance between two fibers, known as fiber collision radius, is 62”. Because fiber collision is more common in over-dense regions, its effect correlates strongly with the underlying clustering. Thus, it is a critical observational systematic that needs to be addressed and mitigated.

In principle, one can overcome this issue by repeatedly visiting the same area of the sky with the telescope, but that significantly reduces the survey efficiency. Thus, a typical survey strategy, such as the ones for BOSS and DESI (Blanton et al. 2003; Abareshi et al. 2022), will only produce spectra for 80-98% of the targets. While the incompleteness is small, it can still produce significant effects on the measured clustering, especially at the precision achievable with DESI (e.g. Pinol et al. 2017; Hahn et al. 2017).

Several techniques have been developed to correct for fiber collision, such as Guo et al. (2012) and Bianchi & Percival (2017). However, these techniques appeal to properties of 2-point correlation function and are not applicable for arbitrary summary statistics.

In principle, one can forward model the effects of fiber collision given the fiber assignment strategy is publicly available. However, fiber assignment codes involve computational expensive steps such as group finding and neighbor searches. These operations make the fiber assignment code prohibitively expensive to apply in every forward model evaluation. Thus, in this section, we demonstrate a novel approach to minimize the effect of fiber collision by applying corrections to the data catalogs.

The basic idea is to infer the redshift of the missing galaxies to the best of our abilities. We assume we have accurate photometric measurement of the angular positions of all the missing galaxies, and we also assume we have a corrected full-shape 2PCF measurement down to very small scales, specifically $\xi(r_p, \pi)$. Both of these assumptions are reasonable for current and upcoming spectroscopic surveys. For each missing galaxy, we identify its closest photometric neighbors up to an arbitrary order (up to the user). Then, we use the corrected 2PCF as the probability distribution of neighboring galaxies around another galaxy, specifically given the transverse distance $r_p$, $\xi(r_p, \pi)$ gives a one dimensional PDF for the LOS distance $\pi$. We can then sample this PDF of LOS distance to probabilistically guess the LOS position of the missing galaxy. We can also fold in the $\pi$ PDFs from additional nearby neighbors to improve the constraints on the missing galaxy’s position. In principle, this idea is similar to “clustering-based redshifts” developed in Ménard et al. (2013), except this general approach is particularly suited for the fiber collision problem because every collided target is necessarily close in projection to a target with known redshift.

To test the performance of this technique, we populate the 25 base simulation cones with an HOD matched to CMASS LRGs, and
apply the full set of selections and masks as described in Section 2.3, from which we measure the desired galaxy clustering statistics and calculate the average over 25 lightcones as the pre-collision “true” measurement. Then we run a BOSS-like tiling and fiber assignment code on the lightcone catalogs to separate galaxies into ones with assigned fibers and ones without. We can then measure the desired clustering statistics on the galaxies with assigned fibers and assess the effect of fiber collision. We refer the readers to Blanton et al. (2003) for a pedagogical description of the tiling and fiber assignment procedure, but offer a brief summary as follows.

The procedure first applies tiling by drawing overlapping circle tiles around a grid of tile centers on the 2D mock sky, with tile center separation and tile radius set to BOSS values. Then within each tiled region, we run a group finder on the projected 2D galaxy field with linking length set to BOSS fiber collision distance 62”.

For each group, we identify the maximum un-collided set, which is the maximum subset of galaxies in the group that are all separated by at least 62”. These galaxies are guaranteed fibers. Then the ones not in the maximum subset are then passed through an additional filtering step, which determines its probability of receiving a fiber based on number of tile overlaps at the location. Specifically, we base these probabilities from BOSS, 0% if there is only one tile, 60% if there are two tiles, 90% if there are three tiles (Reid et al. 2016). As a result, ~ 5% of the galaxies do not receive a fiber, consistent with the collided fraction seen in BOSS CMASS sample (Anderson et al. 2012; Guo et al. 2012; Reid et al. 2016).

Figure 3 showcases the effect of fiber collision on the $k$NN-CDF of the lightcone mocks, averaged over 25 lightcones. The blue curves show the fractional effect of fiber collision. The green bands show the expected sample variance corresponding to the CMASS volume. Clearly, fiber collision has a significant effect on the measured signal. We repeat the same experiment for the projected 2-point correlation function $w_p$ in Figure 4, where the blue curve shows that fiber collision has a significant ~ 5% effect on $w_p$ extending to large scales. The $w_p$ effect is again consistent with those reported for BOSS in Anderson et al. (2012); Guo et al. (2012), validating our fiber collision implementation.

Finally we apply the correction scheme as described to recover the redshifts of the fiber collided galaxies, and measure the summary statistics on the corrected galaxy mocks. On the catalog level, the median absolute recovery error on the LOS coordinate of the collided galaxies is $\Delta z = 8h^{-1}$Mpc. If we sample the corrected position from the probabilities conditioned on the nearest two neighbors, instead of just the nearest neighbor, we get an even better recovery, with a median $z$ error of $\Delta z = 6h^{-1}$Mpc.

More importantly, we assess the performance of the correction on the summary statistics, starting with the $k$NN-CDFs. On Figure 3, the orange curves report the corrected $k$NN-CDFs relative to the truth, where the LOS positions are inferred from just the nearest neighbor. For the $k$NN-CDFs, we can see that the correction significantly reduces the fiber collision effect to within 1$\sigma$ of the “true” measurement. We also find additional improvements to the performance of the correction when including two nearest neighbors instead of just the nearest neighbor. However, given the current level of sample variance, using just the nearest neighbor seems sufficient for the $k$NN-CDFs.

For the 2PCF, we also find excellent recovery of the underlying true signal. The orange curve of Figure 4 shows the performance of the correction on the projected 2-point correlation function. At scales greater than the fiber collision scale $r_p > 0.5h^{-1}$Mpc, the correction almost perfectly recovers the true signal. At smaller scales, the scheme still results in large improvements compared to the uncorrected measurement, but the residual relative to the true signal is still significant. However, this is not a significant issue since scales below 0.4$h^{-1}$Mpc are also systematics dominated and remain largely uninformative for cosmological analysis (Yuan et al. 2022a; Lange et al. 2022). However, the success of the correction scheme on the projected 2PCF on larger scales is expected because, by definition, the projected 2PCF marginalizes over the LOS positions and is thus not strongly sensitive to fiber collision effects. The large-scale effects we see for the blue curve in Figure 4 is coming from the finite LOS integration length $\sigma_{max}$, which we set to 30$h^{-1}$Mpc for this analysis.

To summarize, we have shown through these tests that we can successfully remove the effects of fiber collision with our redshift recovery scheme, at least to the level of precision required for a CMASS analysis on relevant scales. Further improvements to the method are likely needed for a future DESI analysis, where will utilize measurements of significantly higher precision and extending down to much smaller scales. We reserve that discussion for a future paper.

4 HOD RECOVERY ON MOCK GALAXY LIGHTCONES
In this section, we perform a validation test on our lightcone-based forward model by recovering the underlying HOD parameters of a mock galaxy catalog mimicking realistic observations. The purpose is largely to show that parameter inference with lightcone-based full forward models are computational tractable and that such a routine can accurately recover the parameters of interest despite the added layers of model complexity.
4.1 Mock data setup

To construct the target mock galaxy catalog, we start with 20 lightcones at Planck cosmology but with different realizations. In the AbacusSummit suite, these lightcones are generated from the phase 005-024 boxes. The other 5 phases (000-004) are reserved for model evaluations. For each of the 20 lightcones, we apply a fiducial HOD whose baseline parameter values are \( \log M_{\text{cut}} = 12.8, \log M_{1} = 13.9, \sigma = 0.3, \alpha = 1.0, \kappa = 0.3, \) and an completeness parameter \( ic = 0.41. \) Note again that \( \sigma \) and \( \kappa \) will be fix for the analysis. Then we follow the exact steps described in Section 2 and propagate each of the 20 lightcone catalogs through the CMASS redshift-dependent density filter \( n(z), \) the survey window function, and survey masks. Then we measure the desired summary statistics on each of the 20 mocks, and compute the average as the final target statistics.

For \( k \text{NN}(r), \) we use the first 10 orders, \( k = 1, 2, 3, \ldots, 10. \) For each \( k, \) we sample the CDF at 50 linearly spaced scales between \( r_{\text{min}} = 0.1h^{-1}\text{Mpc} \) and \( r_{\text{max}} = 20h^{-1}\text{Mpc}. \) We further remove scales where the CDF is less than 0.1 or greater than 0.9. The blue curves show the full CDF, whereas the orange points mark the “bins” we use as our target data vector.

Figure 5. The full target mock \( k \text{NN-CDF}(r). \) We use \( k = 1, 2, 3, \ldots, 10 \) for this analysis. For each \( k, \) we sample the CDF at 50 linearly spaced scales between \( r_{\text{min}} = 0.1h^{-1}\text{Mpc} \) and \( r_{\text{max}} = 20h^{-1}\text{Mpc}. \) We further remove scales where the CDF is less than 0.1 or greater than 0.9. The blue curves show the full CDF, whereas the orange points mark the “bins” we use as our target data vector.

“binning” scheme in Figure 5, where we show the full peaked CDFs (pCDFs) in blue and the 219 points we retain for our target data vector in orange. The peaked CDF is adopted for visualization purposes and is simply defined as \( \text{pCDF} = \min(\text{CDF}, 1 - \text{CDF}). \) We put quotation marks around the word “binning” to highlight the fact that we are not in fact integrating the CDF into bins, but simply sampling the CDF at a set of scales.

Similar to the \( k \text{NN-CDF}, \) we compute the projected 2PCF over the 20 fully forward modeled lightcone catalogs and compute the average as our mock target data vector. Specifically, we choose 14 logarithmic bins between \( 0.5h^{-1}\text{Mpc} \) and \( 30h^{-1}\text{Mpc} \) in the transverse separation \( r_{p}, \) and \( \pi_{\text{max}} = 30h^{-1}\text{Mpc}. \) We set the smallest projected scale to match the minimum scale at which our redshift recovery method works well. The target projected 2PCF is visualized in Figure 6.

To generate the covariance matrix for likelihood evaluations, we utilize the 1800 AbacusSummit covariance boxes, each of volume \( (500h^{-1}\text{Mpc})^{3}. \) We apply the fiducial HOD to every box and then calculate the summary statistics, without applying the additional layers of systematics. We then calculate the covariance matrix, which we scale down to match the CMASS volume. We note that this covariance matrix likely underestimates the actual uncertainties because it only accounts for sample variance and not any of the systematic effects. However, for the purpose of a mock test, we just need a well...
Figure 6. The target projected 2PCF $w_p$. The $x$ and $y$ axes denote the transverse separation bins $0.5h^{-1}\text{Mpc} < r_p < 30h^{-1}\text{Mpc}$. We plot $r_p w_p$ for visualization. We use the orange dots to highlight the bins used for this analysis.

Figure 7. The mock correlation matrix of the $k$NN-CDF. The $x$ and $y$ axes denote the 219 bins selected across $k = 1, 2, 3, \ldots, 10$, as shown in Figure 5. There is significant off-diagonal power.

Figure 8. The mock correlation matrix of the projected 2PCF $w_p$. The $x$ and $y$ axes denote the transverse separation bins $0.5h^{-1}\text{Mpc} < r_p < 30h^{-1}\text{Mpc}$. There is significant off-diagonal power at larger scales, suggesting that the variance on larger scales is dominated by cosmic variance whereas the smaller scales are dominated by shot noise.

There is significant off-diagonal power. The covariance matrix is closely related to a sum of the counts-in-sphere statistics up to order $k-1$ (Equation 5). The 2PCF shows significantly less off-diagonal power, especially at small transverse scales, where shot noise dominates. At larger transverse scales, sample variance becomes more important and the bins begin to be correlated.

4.2 Likelihood model

To recover the underlying HOD parameters from the summary statistics computed on the target mocks, we utilize the 5 remaining lightcones, phase 000-004. For each model evaluation, we propose a set of HOD parameters from a flat prior, populate the 5 lightcones with the proposed HOD, and then apply the systematics effects, including redshift selection, survey window and masks. Finally, we compute the summary statistics averaged over the 5 lightcones, which we compare with the target summary statistics and calculate likelihoods using the aforementioned covariance matrix. For this analysis, we adopt a Gaussian likelihood function that accounts for both the desired summary statistics and also the average density. Specifically,

$$
\log L = \frac{1}{2} (x_{\text{proposed}} - x_{\text{target}})^T C^{-1} (x_{\text{proposed}} - x_{\text{target}}) \\
+ \frac{1}{2} \left( \frac{\bar{n} - \bar{n}_{\text{target}}}{\sigma_n} \right)^2
$$

where $x$ is the desired summary statistic, $C$ is the covariance matrix, and $\bar{n}$ is the mean number density. $\sigma_n$ is the uncertainty on the measured mean number density. For a CMASS-like sample, we quote $\sigma_n = 5\%$ (Yuan et al. 2022b; Guo et al. 2015). Here we have assumed a Gaussian likelihood, which is known to be the case for the 2PCF. However, for $k$NN-CDF, we test its Gaussianity with the 1800 realizations we have available through the small boxes. Figure 9 shows the distribution of 3 arbitrary $k$NN-CDF bins across the 1800 realizations, and we do not see any significant non-gaussianity.
4.3 Emulator

Typically, to sample an HOD parameter space until convergence, approximately $10^5$ – $10^6$ likelihood evaluations are required. For this analysis, we use the dynesty nested sampler (Speagle & Barbary 2018; Speagle 2020) as it can sample the posterior space more efficiently than an Markov Chain Monte Carlo sampler. However, given that each forward model evaluation takes approximately 1.5 seconds per lightcone on our machine, and we are evaluating 5 lightcones per likelihood call, a 1,000,000 call chain would take more than 80 days. Thus, we adopt an emulator scheme to speed up the likelihood evaluation.

In cosmology, an emulator refers to a scheme where one interpolates sparse likelihood evaluations with a smooth parametrized model, also referred to as the surrogate model or just the emulator. By training such an emulator model, the idea is to replace the expensive likelihood calls with much cheaper emulator model calls, thus enabling a much faster sampling at the cost of introducing additional errors in the model training. Such emulation schemes have become increasingly popular 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 (e.g. Heitmann et al. 2009; Lawrence et al. 2010; Heitmann et al. 2014; Zhai et al. 2019, 2022; Lange et al. 2022; Kobayashi et al. 2022; Yuan et al. 2022a).

For this analysis, we construct a fully connected neural network as our surrogate model, taking in HOD parameters and outputting the fully forward-modeled summary statistics. For the $k$NN-CDF, we adopt a network of 3 layers as our fiducial model, with 200 nodes in each layer and Randomized Leaky Rectified Linear Units (RReLU) activation. We train the network with the Adam optimizer and a mean squared loss function, where we use the diagonal terms of the mock-based covariance matrix as bin weights. For the 2PCF, we find a 2 layer network to work best, with 100 nodes in each layer and RReLU activation. For testing and validation, we set aside 10% of the training sample as the test set and another 10% as the validation set. For training, we follow a mini-batch routine, where the training set is divided into 100 equal batches, which are then passed the optimizer one at a time.

To generate the training set, we follow the hybrid MCMC+emulator approach first implemented in Yuan et al. (2022a). Since we know the target data vector and the covariance matrix, we can directly sample the likelihood surface with a set of MCMC (Markov Chain Monte Carlo) chains. However, instead of running the chains till convergence, we stop the chain once a certain number of likelihood calls has been reached, as limited by the compute time available. We then use these samples generated by the MCMC as the training set for the emulator. Compared to the standard method where the training set is generated with a space-filling sampling of the parameter space, such as a Latin Hypercube, this method allows for a significantly tighter prior region, resulting in higher density of training samples and thus smaller emulator errors. In this approach, we can also think of the emulator step as continuing the MCMC chain, except with a surrogate likelihood model that is orders of magnitudes faster to calculate. However, with a smaller training range, we also need to make sure that the training is robust against biases towards the mean. We note that similar iterative sampling-emulation ideas were also discussed in Pellejero-Ibañez et al. (2020).

To fit the target $k$NN-CDF, we first run an MCMC chain against the target $k$NN-CDF stopped at 100,000 likelihood calls to generate 100,000 training points. Then we impose a likelihood cut to select the 40,000 training points with the highest likelihoods. This sample then undergoes the 80/10/10 training/validation/test split. The resulting training set is then used to train the neural network, and the validation set is used to check for over-fitting during the training. When the training converges, we test the best-fit model on the test set. We present the following test results.
Figure 10 presents the outsample error of $k$NN($r$) as a fraction of the expected error due to sample variance in a CMASS volume. The blue line denotes the median absolute error whereas the shaded region denotes the extent of the 1$\sigma$ region. Clearly, the emulator error is significantly smaller than the expected CMASS sample variance. Averaging over all the tests and bins, we get a representative value of 0.28, which is the mean emulator error as a fraction of the expected sample variance error.

Figure 11 presents the scatter plot of the true values and the emulator-predicted values for a few randomly selected bins. Again, we see that the prediction error is well within the expected CMASS sample variance, which is shown by the blue band. The plot also shows that there is no significant bias towards the mean in the emulator prediction. Thus, we deem the best-fit neural net model to be unbiased and sufficiently accurate to replace the original likelihood calculations without introducing significant additional errors, at least within the training range.

We follow the exact same procedure for the 2PCF. We conduct tests to ensure that the resulting emulator error is subdominant compared to the data error, and that the emulator predictions are not biased towards the mean of the training range. The resulting mean emulator errors as a fraction of the data error are also around 28% for $k$NN($r$) and 30% for the 2PCF. We do not show the figures for the 2PCF for brevity.

4.4 Parameter recovery

Having trained and tested the emulators for the $k$NNs and the 2PCF, we sample the parameter posteriors given the mock data vectors (Figures 5 and Figure 6) and the mock covariance matrices (Figures 7 and Figure 8). For faster sampling, we use the DYNesty nested sampler (Speagle & Barbary 2018; Speagle 2020). We also impose flat priors bounded with an ellipsoid for all parameters. The ellipsoid is constructed as the minimum-volume ellipsoid that envelopes all training points. We initiate each nested sampling chain with 2000 live points and a stopping criterion of $d \log Z < 0.01$, where $Z$ is the evidence. As expected, we achieve excellent fits for both summary statistics, with best-fit $\chi^2/d.o.f < 1$.

We present the resulting 2D marginalized posteriors in Figure 12 and also summarize the 1D marginalized constraints in Table 1. The blue and red contours denote the 1-3$\sigma$ constraints from the $k$NNs and $w_p$, respectively. The black lines show the truth values. The titles above the 1D marginalized posteriors list the 95% or 2$\sigma$ constraints for each parameter. The main conclusion is that our full forward model approach can obtain unbiased recoveries of all model parameters despite the layers of observational systematics. This is significant for analyses with beyond-2PCF statistics on non-linear scales as our full forward approach does not rely on any specific summary statistics. We have demonstrated that a full forward model for any summary statistic that accounts for the full range of observational systematics is computationally viable and should yield unbiased constraints.

Comparing the $k$NN and 2PCF constraints, we see that the $k$NNs derive competitive constraints compared to the 2PCF. Specifically, the $k$NNs yield stronger constraints on $\log M_{\text{cut}}$ while deriving slightly weaker constraints on $\log M_1$. The fact that the $k$NNs derive stronger constraints on $\log M_{\text{cut}}$ shows the promise of $k$NNs in a cosmology analysis. This is because $\log M_{\text{cut}}$ is the only parameter controlling the occupation of the centrals since we have fixed $\sigma$. Given that the satellite fraction is small, the central occupation largely controls the 2-halo term in the clustering and thus the linear bias. Thus, strong constraints on $\log M_{\text{cut}}$ translates to strong constraints on the amplitude of the linear power spectrum and the growth of structure. This is consistent with the Fisher analysis results of Banerjee & Abel (2021a).

We do, however, expect that the 2PCF would be more constraining in the satellite occupation parameters than the $k$NNs. On the one hand, the $k$NNs measure the counts of galaxies around randomly selected query points. In a clustered data set, the randomly selected query points will necessarily mostly sample the under-dense regions more than the over-dense regions. The 2PCF, on the other hand, measure the data-data counts, thus it necessarily samples mostly the over-dense regions, boosting the sensitivity to 1-halo scaling clustering $r < 1h^{-1}$Mpc. And because we expect the 2PCF to be more sensitive to 1-halo scale clustering, we expect it to have stronger constraints on satellite occupation parameters $\log M_1$ and $\sigma$. It is surprising to find the $k$NNs to be competitive even in the satellite occupation parameters. This might be due to the fact that we include higher $k$ orders for $k$NNs but only consider the projected 2PCF in this analysis. In our companion paper Yuan & Abel (2022), we develop variations to the standard $k$NN formulation that exceeds the constraining power of the full-shape 2PCF, even on satellite occupation parameters.

It is also worth noting that this is not a fair comparison since we have not used the $k$NN below $r < 4h^{-1}$Mpc (see Figure 5), whereas we used the 2PCF all the way down to $0.5h^{-1}$Mpc. The cut on the CDF is placed to remove low signal-to-noise points in the $k$NN and is limited by the grid spacing for the query points we used. A denser query point set would increase the signal-to-noise in the $k$NN on small scales but allow us to meaningfully use those scales at the cost of significantly slower likelihood evaluation. In Appendix A, we present an alternative comparison of the two statistics by applying a minimum scale cut of $r_{p, \text{min}} = 3h^{-1}$Mpc on the 2PCF, in which case the 2PCF becomes significantly less constraining than the $k$NNs.

5 DISCUSSIONS

A key limitation of this analysis is that we have only tested the recovery of HOD parameters at fixed cosmology. The goal of utilizing the full information of the smaller scales is to learn both about galaxy–halo connection and the underlying cosmology. In order to enable the cosmology dependence of the full forward model, we need simulation lightcones at non-standard cosmologies. The QuVisor suite provides lightcones with variable cosmologies, but the resolution is not sufficient for modeling small scales (Villaescusa-Navarro et al. 2020). Efforts are currently underway to produce high-fidelity lightcones at variable cosmologies from the periodic boxes of the ABACUS SUMMIT suite (Hadzhiyska et al. 2022b). We reserve the description of these products and the development of cosmology-dependent full forward models for a future paper.

The use of simulation of lightcones also enables direct modeling of redshift-dependence in the bias model, whether it be an HOD($z$) or another galaxy–halo connection model. This is particularly important given the depth of current and upcoming surveys. For example, with DESI, the LRG sample is expected to span redshift range $0.3 < z < 0.8$ whereas the ELG sample is expected to span $0.6 < z < 1.6$ (Zhou et al. 2020; Raichoor et al. 2020). One clearly expects significantly redshift evolution in the galaxy–halo connection throughout these wide ranges. In a standard approach, one can divide the data into several redshift bins, and analyze each bin with a model template constructed on snapshots of the periodic box at or close to the effective redshift within each bin, such approaches suffer from potential biases due to redshift binning and covariances between different redshift bins that might be hard to account for in a combined
analyses. Instead, we propose the use of our lightcone-based forward model approach where we directly parameterize the redshift evolution in the galaxy bias model and compare data to continuous mocks that span the entire redshift range. We would still need to compute the summary statistics in redshift bins, but as long as we make consistent choices between the lightcone and the data, we would not suffer from any biases due to redshift binning. The model covariance between different redshifts are also naturally accounted for. We are actively working on flexible HOD$(_z$) parametrizations and plan on implementation them within the ARAUS HOD framework (Yuan et al. 2021).

While lightcone-based models hold great promises, generating high-precision lightcones of sufficient volume can be a computational challenge. In this analysis, we average over 5 different lightcones (each covering an octant of the sky) at fixed cosmology to achieve the desired volume in the model and ensure the model sample variance is subdominant compared to other sources of uncertainties. With current lightcone implementations such as the one used in Hadzhiyska et al. (2022b), one can only generate one independent lightcone per simulation box. Thus for a cosmology+HOD analysis, one would require repeat simulations per cosmology or a single box that is significantly larger then what is currently available.

However, we might not need repeat lightcones to reduce sample variance after all. At small scales, sample variance is subdominant to other modeling uncertainties and observational systematics, in which case a single lightcone would be sufficient. At larger scales, there are sample variance suppression techniques such as the one developed in Kokron et al. (2022), which uses fast repeat runs of Zel’довich codes to achieve orders of magnitude reduction to sample variance on large scales. In principle, supplementing lightcone-based predictions with this technique should result in sufficiently precise model templates across a wide range scales.

### Table 1. The posterior constraints of the 4 HOD model parameters recovered from the projected 2PCF $w_p$ and the $k$NNs.

| Parameter       | Meaning                                | truth | $w_p$ constraints (95% C.L.) | $k$NN constraints (95% C.L.) |
|-----------------|----------------------------------------|-------|-------------------------------|------------------------------|
| $\log_{10} M_{\text{cut}}$ | The typical mass scale to host a central galaxy | 12.8  | 12.79 ± 0.04                  | 12.80 ± 0.02                 |
| $\log_{10} M_1$  | The typical mass scale for halos to host one satellite | 13.9  | 13.90 ± 0.09                  | 13.91 ± 0.11                 |
| $\alpha$        | The power-law index for satellites      | 1.0   | 1.0 ± 0.2                     | 1.0 ± 0.2                    |
| $\text{ic}$     | The incompleteness parameter           | 0.41  | 0.41 ± 0.03                   | 0.41 ± 0.03                  |

### 6 CONCLUSIONS

In this paper, we constructed and tested a full forward modeling pipeline for galaxy clustering statistics that utilizes high-fidelity simulation lightcones and account for the full range of observational systematics. We demonstrated that one can recover unbiased model constraints using our forward modeling pipeline with the standard 2PCF and the novel $k$-th nearest neighbor statistics (sensitive to correlation functions of all orders) on non-linear scales. While we used two summary statistics as examples, our pipeline is agnostic to the statistics used and we fully expect the technique is broadly applicable.
to other novel summary statistics. This is significant for analyses of upcoming cosmological surveys where the use of novel summary statistics and of non-linear scales are necessary to extract the full information content of the vast datasets.

As a part of the pipeline, we also introduced a novel treatment of fiber collision effects. Specifically, we use the measured clustering to probabilistically recover the redshifts of fiber collided targets. We tested this technique on both the projected 2PCF and the \( k \)NNs and found good redshift recovery compared to other systematics budget. This method is promising as it does not appeal to any specific properties of summary statistics and is in principle applicable to a wide array of beyond-2PCF analyses. We propose additional testing of the method and potential enhancements by leveraging additional information, such as combining with photometric redshift inferences.

By testing our pipeline with two different summary statistics, we also produced a realistic comparison of the \( k \)-th nearest neighbor statistics to the standard 2PCF. We showed that the \( k \)NNs derive stronger constraints on the linear bias, and the two statistics are similarly informative of the properties of satellite galaxies. We explore additional variations to the \( k \)NN formalism in the companion paper Yuan & Abel (2022).

**ACKNOWLEDGEMENTS**

We would like to thank Risa Wechsler, Arka Banerjee, Ashley Ross, Sebastian Wagner-Carena, Philip Mansfield, and others for useful feedback and suggestions in various stages of this analysis.

This work was supported by U.S. Department of Energy through grant DE-SC0013718 and under DE-AC02-76SF00515 to SLAC National Accelerator Laboratory.

This work used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

The Abacussummit simulations were conducted at the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725, through support from projects AST135 and AST145, the latter through the Department of Energy ALCC program.

**DATA AVAILABILITY**

The simulation data are available at https://abacussummit.readthedocs.io/en/latest/. The AbacussHOD code package is
publicly available as a part of the ABACUSUTILS package at http://https://github.com/abacusorg/abacusutils. Example usage can be found at https://abacusutils.readthedocs.io/en/latest/hod.html.

REFERENCES

Abaresi B., et al., 2022, arXiv e-prints, p. arXiv:2205.10939
Alcock C., Paczynski B., 1979, Nature, 281, 358
Anderson L., et al., 2012, MNRAS, 427, 3435
Anderson L., et al., 2014, MNRAS, 441, 24
Banerjee A., Abel T., 2021a, MNRAS, 500, 5479
Banerjee A., Abel T., 2021b, MNRAS, 504, 2911
Bianchi D., Percival W. J., 2017, MNRAS, 472, 1106
Bianchi D., et al., 2018, MNRAS, 481, 2338
Blanton M. R., Lin H., Lupton R. H., Maley F. M., Young N., Zehavi I., Loveday J., 2003, AJ, 125, 2276
Carlson J., White M., Padmanabhan N., 2009, Phys. Rev. D, 80, 043531
Carlson J., Reid B., White M., 2013, MNRAS, 429, 1674
Chapman M. J., et al., 2021, arXiv e-prints, p. arXiv:2106.14961
D’Amico G., Donath Y., Lewandowski M., Senatore L., Zhang P., 2022, arXiv e-prints, p. arXiv:2206.08327
DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00036
Dawson K. S., et al., 2013, AJ, 145, 10
Dawson K. S., et al., 2016, AJ, 151, 44
Eisenstein D. J., et al., 2005, ApJ, 633, 560
Eisenstein D. J., et al., 2018, Phys. Rev. D, 105, 083517
Eisenstein D. J., et al., 2022, arXiv e-prints, p. arXiv:2204.13717
Eisenstein D. J., et al., 2006, MNRAS, 368, 21
Gil-Marín H., Percival W. J., 2019, MNRAS, 485, 3370
Guo H., et al., 2015, MNRAS, 446, 578
Hahn C., Scoccimarro R., Blanton M. R., Tinker J. L., Rodríguez-Torres S. A., 2017, MNRAS, 467, 1940
Heitmann K., Higdon D., White M., Habib S., Williams B. J., Lawrence E., Wagner C., 2009, ApJ, 705, 156
Heitmann K., Lawrence E., Kwan J., Habib S., Higdon D., 2014, ApJ, 780, 111
Kobayashi Y., Nishimichi T., Takada M., Miyatake H., 2022, Phys. Rev. D, 105, 083517
Leauthaud A., 2014, MNRAS, 446, 578
Loveday J., 2003, AJ, 125, 2276
Loveday J., et al., 2020, Research Notes of the American Astronomical Society, 4, 181
MNRAS 000, 1–15 (2022)

APPENDIX A: COMPARING THE STATISTICS ON EQUIVALENT SCALES

Figure 12 appears to show the 2PCF as more informative on the HOD than kNNs. However, as we pointed out towards the end of section 4.4, we have only utilized the kNNs on scales above \( r > 4 h^{-1} \text{Mpc} \) due to the relatively dense query set, whereas we used the 2PCF all the way down to \( 0.5 h^{-1} \text{Mpc} \). Here we facilitate a comparison of the two statistics at equivalent scales, specifically by only using the 2PCF at scales \( r_p > 3 h^{-1} \text{Mpc} \). We follow the exact same procedure as laid out in section 4 and show the resulting constraints in Figure A1.

We see that the constraining power of the projected 2PCF \( w_p \) decreases considerably in both the central and satellite HOD parameters. The constraints on the satellite parameter \( \alpha \) turns out particularly poor, hitting prior bounds in both directions. The \( w_p \) at \( r_p > 3 h^{-1} \text{Mpc} \) still derives strong constraints on the mass parameters \( \log M_{\text{cut}} \) and \( M_f \), but the constraining power is now significantly inferior to that of the kNN.

Finally, we note that the definition of scales is different between the 2PCF and kNNs, with the 2PCF defining the separation between data-query pairs and kNNs defining the separation between data-query pairs. The idea of testing on equivalent scales is an attempt to compare the statistics in roughly the same 2-halo regime, but an exactly fair comparison is not possible. We also point to the fact that there are significantly more bins available in the kNN than the projected 2PCF.
Figure A1. The HOD posterior as recovered by the projected 2PCF $w_p$ with $r_{p, \text{min}} = 3h^{-1}\text{Mpc}$ and the $k\text{NN}(r)$. The black lines denote the truth values, whereas the contours denote the 1-3$\sigma$ constraints. The titles on the 1D histograms describe 95% confidence interval of the $k\text{NN}$ constraints.

Ultimately, these results highlight the complementarity of the two statistics and the potential information gain by combining the two.

This paper has been typeset from a \LaTeX\ file prepared by the author.