PHYSICAL REVIEW D **93,** 063512 (2016)

Towards a self-consistent halo model for the nonlinear large-scale structure

Fabian Schmidt

Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Strasse 1, 85748 Garching, Germany

(Received 22 November 2015; published 11 March 2016)

The halo model is a theoretically and empirically well-motivated framework for predicting the statistics of the nonlinear matter distribution in the Universe. However, current incarnations of the halo model suffer from two major deficiencies: (i) they do not enforce the stress-energy conservation of matter; (ii) they are not guaranteed to recover exact perturbation theory results on large scales. Here, we provide a formulation of the halo model (EHM) that remedies both drawbacks in a consistent way, while attempting to maintain the predictivity of the approach. In the formulation presented here, mass and momentum conservation are guaranteed on large scales, and results of the perturbation theory and the effective field theory can, in principle, be matched to any desired order on large scales. We find that a key ingredient in the halo model power spectrum is the halo stochasticity covariance, which has been studied to a much lesser extent than other ingredients such as mass function, bias, and profiles of halos. As written here, this approach still does not describe the transition regime between perturbation theory and halo scales realistically, which is left as an open problem. We also show explicitly that, when implemented consistently, halo model predictions do not depend on any properties of low-mass halos that are smaller than the scales of interest.

DOI: 10.1103/PhysRevD.93.063512

I. INTRODUCTION

In the halo model (see [1] for a review), all matter in the Universe is assumed to be within virialized structures, called halos. Under this assumption, the statistics of matter on all scales are determined by the statistics of these halos as well as their density profiles. Most incarnations of the halo model further assume that halos are mutually exclusive, such that each mass element is part of one and only one halo, and we will do so as well here.

Currently, frequently employed incarnations of the halo model (e.g. [2–21]) have two major deficiencies: (i) they do not enforce the physical constraint of stress-energy conservation of matter; (ii) they are not guaranteed to recover exact perturbation theory (PT) results on large scales. The most widely known symptom of these deficiencies is the k-independent white noise contribution to the matter power spectrum $P_{mm}(k)$ of the 1-halo term on large scales. The goal of this paper is to address these issues while attempting to preserve the successes of the halo model, namely its predictivity: the ability to provide a reasonably good description of matter and halo statistics over a wide range of scales with few free parameters.

For this reason, we also demand (iii) *self-consistency*, namely that the same set of parameters describes the nonlinear n-point correlations of matter as well as the cross-correlations of matter with halos. This crucially requires that nonlinear and nonlocal bias is incorporated in the model. Further, we demand that the halo model also consistently describe the cross-correlation of the nonlinear matter density with the initial conditions, i.e. in case of the power spectrum, the matter power spectrum $P_{mm}(k)$ and the cross-correlation (propagator) between the initial density field evolved forward using PT, and the final, nonlinear density fields, $P_{1m}(k)$. The former receives stochastic contributions, while the latter only contains the *deterministic* terms, i.e. contributions that correlate with initial perturbations with wave numbers of the order of k and smaller. $P_{mm}(k)$ and $P_{1m}(k)$ can be used to extract the stochastic contribution to the matter power spectrum in simulations [22,23].

To this end, we describe how the halo model can be constructed consistently up to a given order, with a finite set of free parameters, so that it is guaranteed to satisfy mass and momentum conservation on scales much larger than individual halos, as well as matching the exact perturbative solution to the same order on large scales, including effective beyond-fluid terms (but see next paragraph). In this sense, the halo model consistently extends the predictions of the effective field theory (EFT) of large-scale structure (LSS) [24] into the nonlinear regime. Of course, on fully nonlinear scales the predictions are not guaranteed to be correct or within a rigorously calculable theoretical uncertainty of the correct answer. Since various implementations of the halo model paradigm have been presented in the literature, we will adopt the shorthand EHM for the specific construction presented here.

There is a further well-known trouble with the halo model which we *do not* address: in the transition region between large scales where perturbation theory is valid, and small scales that are mostly determined by halo density profiles (1-halo regime), the halo model is known to not describe simulation results well. Moreover, the situation becomes worse when going to higher order in PT. As the focus of this paper is on a consistent description of large scales, we will not have much to say on this here. It is likely however that the halo model implementation presented here will need to be extended to solve this issue (see Sec. IV).

Let us briefly describe the relation to previous attempts at resolving the above mentioned issues of the halo model. Note that these attempts were partially motivated by modeling the transition regime mentioned above. References [25,26] (Halo-PT) performed a separation of matter statistics into n-halo terms in Lagrangian space. This offers the advantage of a simple implementation of mass conservation. On the other hand, one needs to assume a specific exclusion model for halos, and it is not possible to specify the bias parameters of halos in the model. Thus, a consistent connection to perturbation theory on large scales does not appear to be feasible in this approach. Further, as pointed out in [25], the stochastic contribution to $P_{mm}(k)$ does not scale as k^4 in the low-k limit, as required by mass and momentum conservation, but as k^2.

References [27,28] give a prescription (Halo-Zel'dovich) for the matter power spectrum and its covariance based on the power spectrum in the Zel'dovich approximation, to which a power series in k^2 is added. The latter can be interpreted as an expansion of the mean halo profile. At low k, the lowest order coefficient can be matched to the 1-loop power spectrum predicted by perturbation theory, in order to achieve the correct large-scale limit. For this, Ref. [28] had to introduce a compensating kernel in order to cancel the k^0 contribution from the profile expansion. Note that the stochastic and deterministic contributions to $P_{mm}(k)$ are not modeled separately in this approach, which only considers their sum. Since the model is built on the matter power spectrum in the Zel'dovich approximation, nonlinear halo bias is not included in this prescription.

Thus, while these ansatzes recognize the problems of the standard halo model, and point to possible approaches to solve the transition regime problem, they do not satisfy all of the conditions (i)–(iii) mentioned above in their current form, since they do not consistently describe halo correlations and the stochastic contribution to $P_{mm}(k)$.

The outline of the paper is as follows. Section II describes the general procedure for constructing EHM, and spells out the assumptions made; this section constitutes the core of the paper. We then describe the lowest order (tree-level) incarnation of EHM and its prediction for the matter power spectrum in Sec. III. Section IV discusses aspects of the next-to-leading order (1-loop) EHM power spectrum prediction. After that, we consider the tree-level bispectrum in Sec. V. Section VI contains a brief discussion of the matter velocity field. We make some comments regarding the relation to the EFT of LSS in Sec. VII, before concluding in Sec. VIII. The appendixes discuss the issue of the low-mass cutoff and halo triaxiality, and present the expressions for the bispectrum that are too lengthy for the main text. The discussion of the low-mass cutoff in Appendix A is also relevant for other frequently used versions of the halo model.

II. GENERAL SELF-CONSISTENT HALO MODEL

In this section we describe the general procedure for relating the matter density perturbation,

$$\delta_m(\mathbf{x},\tau) \equiv \frac{\rho(\mathbf{x},\tau)}{\bar{\rho}(\tau)} - 1, \tag{1}$$

where $\bar{\rho}$ is the background density, to halo properties. We begin by allowing for fully general halo clustering and profiles. Afterwards, we assemble δ_m and show what constraints mass and momentum conservation of matter place on the halo properties.

A. Halo clustering

Let us begin with the description of the halo density field at fixed mass M. In slight abuse of notation, we denote the local number density of halos per logarithmic mass interval as $n(M,\mathbf{x},\tau)$. This will not lead to confusion as we will never consider any other type of halo mass function. The cosmological average of the same quantity is defined as $\bar{n}(M,\tau)$. The number density perturbation of halos at a given mass is correspondingly denoted as

$$\delta_{h,M}(\mathbf{x},\tau) \equiv \frac{n(M,\mathbf{x},\tau)}{\bar{n}(M,\tau)} - 1. \tag{2}$$

Let us consider large scales, that is, scales much larger than the Lagrangian radius $R_L(M)$. The EHM ansatz we will pursue here assumes that all higher derivative terms are supplied by halo profiles. Then, it is sufficient to describe the clustering of halos at lowest order in derivatives, significantly reducing the number of free parameters of the model. Relaxing this assumption is one possibility to address the failures of EHM in the transition regime (Sec. IV) however. The equivalence principle guarantees the absence of halo velocity bias at lowest order in derivatives [29]. In other words, at lowest order in derivatives halos move along the trajectories of the matter fluid itself. We can then write, to any given order in perturbation theory,

$$\delta_{h,M}(\mathbf{x},\tau) = \sum_O \{b_O(M,\tau) + \epsilon_O(M,\mathbf{x},\tau)\}[O](\mathbf{x},\tau) + [\epsilon](M,\mathbf{x},\tau), \tag{3}$$

where $b_O(M,\tau)$ are bias parameters, and $[O]$ are renormalized bias operators constructed out of the density, tidal field and convective time derivatives of the same.[1] Complete bases for the bias expansion have been described in [29,30]; renormalization of bias operators is described in [31,32]. The fields ϵ, ϵ_O are stochastic fields with zero means which are completely characterized by their moments $\langle[\epsilon](M,\mathbf{x},\tau)[\epsilon](M',\mathbf{x},\tau)\rangle$ and so on (again, this holds at lowest order in derivatives). The explicit bias expansion to linear order [Eq. (22)] and second order

[1]Note that it would be more accurate to write $[\epsilon_O O]$, since this combination is renormalized jointly.

[Eq. (46)] will be given below. In general, the bias parameters of halos are not uniquely determined by their mass, a phenomenon known as assembly bias. We neglect this effect in the main text and discuss it briefly in Sec. VIII.

B. Halo profiles

Further, we also need a prescription for the density profiles of halos, which we write as

$$\rho(\mathbf{r}, M, \tau) = M y(\mathbf{r}, M, \tau). \tag{4}$$

We enforce the following mass constraint for the profile:

$$\int \rho(\mathbf{r}, M, \tau) d^3\mathbf{r} = M \int y(\mathbf{r}, M, \tau) d^3\mathbf{r} = M. \tag{5}$$

In EHM, this constraint is essential in order for Eq. (3) to be consistent, and for exact perturbation theory to be matched on large scales. It states that the mass function and bias parameters defined in Sec. II A completely characterize the mass distribution on large scales (at lowest order in derivatives), while the halo profiles provide the detailed distribution on small scales. We denote the Fourier transform of $y(\mathbf{r}, M, \tau)$ (which is dimensionless) as $y(\mathbf{k}, M, \tau)$.

Let us assume a mean spherically averaged profile $y(r, M, \tau)$ (we generalize this below). Besides the mass, the halo profile, averaged over an ensemble of halos within a finite region, also depends on the local density and tidal field. In general, we should perturbatively expand the profiles in the local gravitational observables in the same way as the halo abundance [Eq. (3)], where now the bias parameters and stochastic fields become functions of r as well. This formidable set of free functions can however be reduced by using the fact that spherically averaged halo profiles are usually well described by a single number (apart from the mass), for example, in case of the NFW profile [33], the concentration c. Then, it is sufficient to write $y = y(r, M, \tau, c)$ and expand the concentration in a bias expansion of the same type as in Eq. (3),

$$\frac{c(M, \mathbf{x}, \tau)}{\bar{c}(M, \tau)} = 1 + \sum_O \{b_O^c(M, \tau) + \epsilon_O^c(M, \mathbf{x}, \tau)\}[O](\mathbf{x}, \tau) + [\epsilon^c](M, \mathbf{x}, \tau), \tag{6}$$

where $\bar{c}(M, \tau)$ denotes the mean halo concentration. Assuming that the fractional fluctuations in the concentration are much less than one, we can then expand

$$y(r, M, \tau, c(\mathbf{x}, \tau)) = y(r, M, \tau, \bar{c}) + y_c(r, M, \tau, \bar{c})[b_1^c(M, \tau)\delta(\mathbf{x}, \tau) + [\epsilon^c](M, \mathbf{x}, \tau) + \cdots] \tag{7}$$

where

$$y_c(r, M, \tau, c) \equiv \frac{\partial}{\partial \ln c} y(r, M, \tau, c). \tag{8}$$

Note that Eq. (5) implies that $\int d^3\mathbf{r} y_c(r, M, \tau, c) = 0$.

In general, we should also take into account that halos are triaxial. This was investigated in [34]. Moreover, the orientation of the axes will correlate with large-scale tidal fields. We study this in Appendix B, and find that, under reasonable assumptions, the terms introduced by allowing for halo triaxiality are degenerate with those obtained through the isotropic profile expansion Eq. (7). Thus, we can effectively account for triaxiality through this expansion. This is very useful as it reduces the number of free parameters in the halo model predictions.

Let us consider the Fourier transform of the profile on large scales, i.e. at low k. Equation (5) implies that $y(k \to 0, M, \tau) = 1$. Moreover, we can expand

$$y(k, M, \tau, c) \overset{k\to 0}{=} 1 - a_M k^2 R_M^2 + \mathcal{O}(k^4), \tag{9}$$

where R_M is the Eulerian halo radius (e.g. R_{200}) and a_M is a mildly mass-dependent number of the order of one that depends on the exact profile and mass-concentration relation assumed. Equation (9) will be useful when considering the low-k limit of matter statistics in the halo model. Further, we immediately see that

$$y_c(k, M, \tau, c) \overset{k\to 0}{=} -\frac{\partial a_M}{\partial \ln c} k^2 R_M^2 + \mathcal{O}(k^4), \tag{10}$$

scaling as k^2 in the $k \to 0$ limit.

C. Matter density

Following the halo model paradigm, the matter density perturbation δ_m is given by a superposition of halos weighted by their density profiles. Let us denote the frequently appearing mass-weighting integral as

$$\int d\rho(M, \tau) \equiv \int d\ln M \frac{M}{\bar{\rho}} \bar{n}(M, \tau). \tag{11}$$

Note that $d\rho$ is dimensionless, and that $\int d\rho(M, \tau) = 1$ in order to satisfy *global* mass conservation at the background level, which corresponds to the well-known integral constraint on the mass function.[2] Equation (11) formally requires a parametrization $\bar{n}(M, \tau)$ for all masses. We discuss this issue at the end of this section.

The fractional matter density perturbation is then, in full generality, given by

[2]Note that we do not need to assume that the mass function is universal, i.e. determined by a function $f[\delta_c/\sigma(M)]$.

$$1+\delta_m(\mathbf{x}) = \int d\rho(M) \int d^3\mathbf{y}[1+\delta_{h,M}(\mathbf{y})] \times y[\mathbf{x}-\mathbf{y}, M, c(\mathbf{y})], \quad (12)$$

where here and in the following we drop the explicit time argument for clarity (in the following, we always work at some fixed time τ).

Equation (12) by itself is not sufficient however, since $\delta_{h,M}$ in turn is constructed from δ_m. Here, we introduce the following procedure. First, one expands $\delta_{h,M}$ and $c(\mathbf{y})$ to a fixed order in perturbation theory. For example, to match PT predictions for the 1-loop power spectrum, we need to expand $\delta_{h,M}$ to third order in perturbations (Sec. IV). For consistency, one should similarly expand c (around $\bar{c}$) to third order, unless those terms are numerically suppressed (see Sec. III). As we noted above, Eq. (5) ensures that the terms involving the concentration are higher order in derivatives. Then, the desired statistics of δ_m are given as convolutions of correlators of the renormalized operators appearing in the bias expansion Eq. (3) [and Eq. (6)] with the halo density profiles. We see explicit examples of this in the following sections. This approach assumes that all corrections to the PT matter density field are effectively modeled by the halo profiles $y(\mathbf{y}, M, c)$. We discuss the issues related to this assumption in Sec. IV.

It is important to emphasize again that *all* matter and halo statistics follow unambiguously from this procedure, so that the same set of halo properties $\bar{n}(M)$, $b_O(M)$, $y(k, M)$ describe all these observables. Further, in most studies to date, halo model statistics were derived at tree level in perturbation theory. In EHM, this is not necessary, and the halo model can be extended to match perturbation theory at any desired order. We see one example (and the associated issues) in Sec. IV.

The bias expansion Eq. (3) describes the distribution of matter among halos on large scales, i.e. scales much larger than typical sizes of halos, independently of their internal structure. Combining this with the fact that halos are comoving with matter on large scales, it is easy to see that *local* mass and momentum conservation simply imply the following constraints on the bias parameters and stochasticities:

$$\int d\rho(M) b_O(M) = \begin{cases} 1 & O=\delta \\ 0 & \text{otherwise} \end{cases}$$
$$\int d\rho(M)[\epsilon](M,\mathbf{k}) \overset{k\to 0}{=} 0 + \mathcal{O}(k^2)$$
$$\int d\rho(M)[\epsilon_O(M)O](\mathbf{k}) \overset{k\to 0}{=} 0 + \mathcal{O}(k^2), \quad (13)$$

which hold at all times. The first line states that the mass-weighted mean linear bias of halos should be 1, while the corresponding mean bias vanishes for all nonlinear terms. The conditions on ϵ, ϵ_O are to be understood as constraints on the auto and cross correlations between the renormalized stochastic fields in the low-k limit. That is, they imply for example

$$\int d\rho(M) \int d\rho(M') \langle [\epsilon_O(M)O](\mathbf{k})[\epsilon_{O'}(M')O'](\mathbf{k}')\rangle' \overset{k\to 0}{=} \mathcal{O}(k^4). \quad (14)$$

One can also interpret the stochasticity constraints locally however: if we consider the matter density at a given point, coarse grained on a sufficiently large scale (much larger than the radius of typical halos), then the stochasticity of halos of various mass cancels after mass weighting. That is, there might be more halos at some fixed mass in a given realization of initial phases, but this has to be compensated by a smaller number of halos at other masses such that the total amount of matter is locally conserved.

Apart from ensuring mass and momentum conservation, these conditions are sufficient to ensure that on scales larger than halos, the matter density Eq. (12) reduces to the perturbation theory prediction at the desired order. The constraint on the stochasticity will become particularly relevant in the following sections, as it is responsible for removing the constant tail of the standard 1-halo term in the low-k limit. Naturally, any constraints that can be placed on the opposite, small-scale limit are very useful as anchor points. First, most obviously, one can use the existing, very accurate measurements of mean halo profiles. Second, we can also place physical constraints on the stochastic terms in the high-k limit.

Before turning to this limit, let us discuss another issue related to Eq. (13). The integral $\int d\rho(M)$ formally extends to arbitrarily small halo masses, far beyond the range that is empirically calibrated with simulations. In fact, for standard parametrizations of the mass function, the mass-weighting integrals in Eq. (13) typically converge very slowly towards low masses. Since properties of very low-mass halos are poorly constrained by simulations, this raises the question of whether the halo model predictions discussed here and presented in the literature actually rely on extremely low-mass halos whose properties are poorly known.

Fortunately, as we show in Appendix A, the answer is no. Specifically, if the properties of halos are calibrated to a minimum mass M_s, then one can cut off the mass-weighting integral below M_s, and introduce compensating parameters to enforce the conditions in Eq. (13). After this procedure, any systematic uncertainties in the halo model predictions due to the mass cut scale as $(kR_{M_s})^2$. These systematics reach 10% at a scale of

$$k_{10\%} \approx 5.6h\,\mathrm{Mpc}^{-1}\left(\frac{M_s}{10^{10}h^{-1}M_\odot}\right)^{-1/3}. \quad (15)$$

Given the current advanced state of high-resolution simulations, this is not likely to be an important constraint for cosmological applications of the halo model. Note that the

procedure we describe in Appendix A applies to any halo model prescription that involves integrals over halo masses.

D. Stochasticity in the high-k limit

The stochastic terms ϵ, ϵ_O are nonperturbative and numerically important in the high-k limit. For scales much smaller than the sizes of halos (of a given mass), the stochasticity in the halo abundance should approach Poisson statistics governed by the *local* halo abundance $\bar{n}(M)[1+\delta_{h,M}]$. This is because Poisson statistics apply if the wavelength $1/k$ of a given mode is much smaller than the mean interhalo separation, that is, if $\bar{n}/k^3 \ll 1$. Further, since halos are nonoverlapping in the halo model (each matter particle only belongs to one parent halo), halos of different mass have independent Poisson noise. These are significant constraints, since in this limit, they completely determine the moments of $[\epsilon]$ as well as all $[\epsilon_O]$. This works as follows. For clarity, we drop the brackets around ϵ, ϵ_O in the remainder of this section, keeping in mind that we always deal with the renormalized fields.

Consider halos within an infinitesimal logarithmic mass interval $d\ln M$ centered around a fixed mass M, and a fictitious small volume element V around point $\mathbf{x}$ such that

$$\bar{N} \equiv V\bar{n}d\ln M \ll 1. \tag{16}$$

The Poisson assumption states that the halo number within this volume follows a Poisson distribution,

$$N(\mathbf{x}) \sim \text{Poisson}\left[\bar{N}\left(1+\sum_O b_O[O](\mathbf{x})\right)\right]. \tag{17}$$

Here, the operators $[O](\mathbf{x})$ are considered to be coarse grained on some larger scale (of the order of the halo radius, for example). We can subtract the mean, which corresponds to the deterministic part of the bias expansion Eq. (3), and call the remainder $\bar{N}\epsilon_p(\mathbf{x})$ with $\langle\epsilon_p\rangle = 0$. Equation (17) then specifies the moments of ϵ_p, i.e.

$$\begin{aligned}\langle\epsilon_p^2\rangle &= \frac{1}{\bar{N}}\left(1+\sum_O b_O[O](\mathbf{x})\right)\\ \langle\epsilon_p^3\rangle &= \frac{1}{\bar{N}^2}\left(1+\sum_O b_O[O](\mathbf{x})\right),\end{aligned} \tag{18}$$

and so on. On the other hand, we have a specific perturbative expansion of the stochasticity in Eq. (3), which yields

$$\epsilon_p(\mathbf{x}) = \epsilon(\mathbf{x}) + \sum_O [\epsilon_O O](\mathbf{x}). \tag{19}$$

By matching Eq. (19) to the moments derived from Eq. (18), and using the fact that there is only a single random field ϵ_p (at fixed halo mass), we can then uniquely determine the moments of ϵ and ϵ_O, order by order. Performing a Fourier transform within the volume V, we then obtain the desired high-k limit of the moments in Fourier space. For example, at linear order we simply have

$$\langle\epsilon(M,\mathbf{k})\epsilon(M',\mathbf{k}')\rangle' \overset{k\to\infty}{=} \frac{\delta_D(\ln M - \ln M')}{\bar{n}(M)}, \tag{20}$$

where a prime denotes that the momentum conserving delta function has been dropped. At second order, we obtain the following two additional constraints:

$$\begin{aligned}&\langle\epsilon(M,\mathbf{k})\epsilon(M',\mathbf{k}')\epsilon(M'',\mathbf{k}'')\rangle'\\ &\quad\overset{k\to\infty}{=} \frac{\delta_D(\ln M - \ln M')\delta_D(\ln M - \ln M'')}{[\bar{n}(M)]^2}\\ &\langle\epsilon(M,\mathbf{k})\epsilon_\delta(M',\mathbf{k}')\rangle'\\ &\quad\overset{k\to\infty}{=} \frac{1}{2}b_1(M)\frac{\delta_D(\ln M - \ln M')}{\bar{n}(M)}.\end{aligned} \tag{21}$$

These are all stochastic moments that exist at second order (Sec. V). The second line of Eq. (21) is directly related to the halo sample variance discussed in [10]. Note that both Eqs. (20) and (21) violate the constraints Eq. (13) in the opposite, large-scale limit. This already shows that a scale-dependent stochasticity is a necessary part of a consistent formulation of the halo model. The entire reasoning of this section also applies to the stochastic fields appearing in the profile expansion Eq. (6). Moreover, in the high-k limit these fields are uncorrelated with the stochasticity in the halo number.

III. LOWEST-ORDER HALO MODEL AND POWER SPECTRUM

The lowest-order consistent incarnation of the halo model expands Eq. (3) to linear order,

$$\delta_{h,M}(\mathbf{x}) = b_1(M)\delta_1(\mathbf{x}) + [\epsilon](M,\mathbf{x}), \tag{22}$$

where δ_1 denotes the linear density field. In addition, the profiles are expanded via Eq. (7),

$$\begin{aligned}y(r,M,c(\mathbf{x})) &= y(r,M,\bar{c})\\ &\quad + [b_1^c(M)\delta_1(\mathbf{x}) + [\epsilon^c](M,\mathbf{x})]y_c(r,M,\bar{c}).\end{aligned} \tag{23}$$

The matter density perturbation is then given as a mass-weighted integral of the halo number density convolved with the halo density profile as in Eq. (12).

Let us first look at the matter propagator, i.e. the cross-correlation of δ_m with the PT-evolved density field (here just the linear density field) in Fourier space,

$$P_{1m}(k) = \int d\rho(M)[b_1(M)y(k,M) + b_1^c(M)y_c(k,M)] \times P_{\rm L}(k), \quad (24)$$

where here and in the following we drop the explicit concentration argument when it is set to the mean value $\bar{c}(M)$, and $P_{\rm L}$ denotes the linear matter power spectrum. Using the low-k behavior of y and y_c and Eq. (13), we see that in the low-k limit we recover

$$P_{1m}(k) = P_{\rm L}(k)[1 + \mathcal{O}(R_{\rm HM}^2 k^2)], \quad (25)$$

where

$$R_{\rm HM}^2 \equiv \int d\rho(M) a_M b_1(M) R_M^2. \quad (26)$$

This is the characteristic scale that appears in the low-k limit of EHM, and it is of the order of R_{M_*}, where M_* is defined through $\sigma(M_*) = \delta_c$; that is, $R_{\rm HM}$ is of the order of the typical Eulerian halo radius. Note that this scale is smaller than the nonlinear scale $1/k_{\rm NL}$ where the density contrast becomes of the order of 1.

We now turn to the matter power spectrum. This is given by

$$\begin{aligned} P_{mm}(k) = \int d\rho(M) \int d\rho(M') \{ & y(k,M)y(k,M')[b_1(M)b_1(M')P_{\rm L}(k) + P^{\epsilon\epsilon}_{MM'}(k)] \\ & + 2y(k,M)y_c(k,M')[b_1(M)b_1^c(M')P_{\rm L}(k) + P^{\epsilon\epsilon^c}_{MM'}(k)] \\ & + y_c(k,M)y_c(k,M')[b_1^c(M)b_1^c(M')P_{\rm L}(k) + P^{\epsilon^c\epsilon^c}_{MM'}(k)]\}, \end{aligned} \quad (27)$$

where we have defined

$$P^{\epsilon^a\epsilon^b}_{MM'}(k) \equiv \langle [\epsilon^a](M,\mathbf{k})[\epsilon^b](M',\mathbf{k}')\rangle'. \quad (28)$$

Equations (24) and (27) differ from the standard halo model power spectrum in two respects: the stochasticity covariances $P^{\epsilon^a\epsilon^b}_{MM'}(k)$ and the terms from the expansion of halo concentration in long-wavelength perturbations, proportional to y_c. We examine both of them in the following sections.

First, however, we consider the low-k limit of Eq. (27). The constraints Eq. (13) imply that

$$\begin{aligned} \int d\rho(M) P^{\epsilon}_{MM'}(k \to 0) &= 0 + \mathcal{O}(k^2) \\ \int d\rho(M) \int d\rho(M') P^{\epsilon}_{MM'}(k \to 0) &= 0 + \mathcal{O}(k^4). \end{aligned} \quad (29)$$

For the cross-correlation between ϵ, ϵ^c on the other hand, we only demand $\int d\rho(M) \int d\rho(M') P^{\epsilon\epsilon^c}_{MM'} = \mathcal{O}(k^2)$. The first line here says that in the low-k limit, the halo stochasticity covariance has (at least) one zero eigenvalue, with the corresponding eigenvector given by mass weighting (see also [35–37]). Reference [36] performed a detailed analysis in simulations. Indeed, they find that the lowest eigenvalue of $P^{\epsilon\epsilon}_{MM'}$ is significantly lower than the shot noise $1/\bar{n}(M)$ of halos in the mass range they considered. Further, the corresponding eigenvector is close to mass weighting. Similar results were found in [38].

Using that $y(k,M) \to 1$ for $k \to 0$, we then see that $P_{mm}(k)$ has the same low-k behavior Eq. (25) as $P_{1m}(k)$. Moreover, the stochastic contributions, i.e. all terms that involve $P^{\epsilon\tilde{\epsilon}}_{MM'}$, scale as k^4 in the low-k limit, just as demanded by mass and momentum conservation. The leading contribution to $P_{mm}(k)$ is then

$$\begin{aligned} P_{mm}(k) = & P_{\rm L}(k) + \mathcal{O}(R_{\rm HM}^2 k^2)P_{\rm L}(k) \\ & + \mathcal{O}[k^4, k^4 P_{\rm L}(k)]. \end{aligned} \quad (30)$$

Note that the 1-loop matter power spectrum contributes terms that also scale as $k^2 P_{\rm L}(k)$, but involve $1/k_{\rm NL}$ instead of $R_{\rm HM}$. This shows that one needs to carry out the "halo model at 1-loop," by extending Eq. (22) to third order, in order to obtain a consistent matching to beyond-perfect-fluid terms in the EFT.

Figure 1 (red solid line) shows the deterministic contribution from the expansion of the halo density, i.e. the first term in the second line of Eq. (27). This is the standard 2-halo term. Given the discussion in the previous paragraph, we do not expect this to be a good match to the true nonlinear power spectrum from simulations. For our numerical results, we assume a flat ΛCDM cosmology with cosmological parameters given by $\Omega_m = 0.27$, $h = 0.7$, $\Omega_b h^2 = 0.023$, $n_s = 0.95$, $\sigma_8 = 0.791$. We use the Sheth-Tormen mass function [39] and associated linear bias, and the concentration-mass relation of [40]. We assume that halo masses are given in terms of a mean interior density equal to the virial density $\rho_{\rm vir} = 363\bar{\rho}$ for this cosmology. All results are shown for $z = 0$.

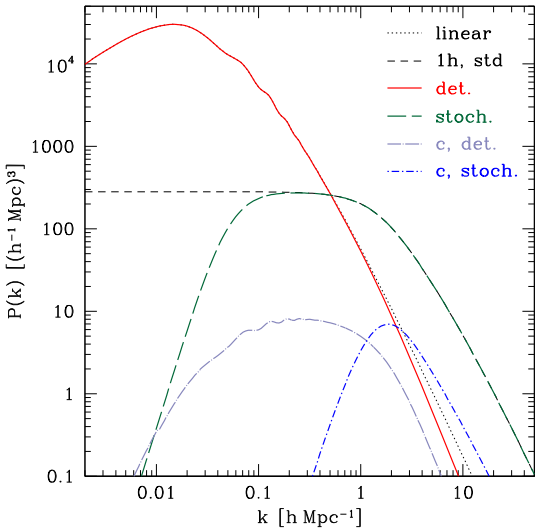

FIG. 1. Contributions to the lowest-order halo model matter power spectrum Eq. (27) at $z = 0$. The red solid line shows the deterministic contribution [first term in the second line of Eq. (27)], i.e. the standard 2-halo term, while the green long-dashed line is the stochastic contribution (second term in the same line). For comparison, we also show the standard 1-halo term as a black thin short-dashed line. The light blue dot–long-dashed line shows the deterministic contributions from the concentration expansion [first terms in the third and fourth lines of Eq. (27)]. Finally, the stochastic terms of the concentration expansion [second terms in the same lines of Eq. (27)] are shown as a blue dot-dashed line. The linear power spectrum is shown as a thin dotted line.

A. Halo stochasticity

The last term in the first line of Eq. (27) is the halo model prediction for the stochastic part of the matter power spectrum. Here, by stochastic we mean that it does not correlate with the initial conditions on the scale k (or at larger scales). In the halo model, this is controlled by the halo stochasticity covariance $P^{\epsilon\epsilon}_{MM'}(k)$. Clearly, this is a key ingredient of the halo model, as important though much less well studied than the mass function, linear bias, and halo profiles. The standard halo model assumes a k-independent diagonal covariance following Poisson statistics,

$$P^{\epsilon\epsilon,\mathrm{std}}_{MM'} = \frac{\delta_D(\ln M - \ln M')}{\bar{n}(M)}, \tag{31}$$

which, following Sec. II D, is only justified in the high-k limit, i.e. well within halos. Moreover, it clearly does not satisfy Eq. (29) and thus violates mass and momentum conservation. Thus, we need to come up with a more physical parametrization of $P^{\epsilon\epsilon}_{MM'}$ at low k, which asymptotes to Eq. (31) if k is larger than the mean interhalo separation, which is directly related to the Lagrangian radii $R_L(M)$, $R_L(M')$, respectively.

One possibility is to simply subtract the trace to ensure a zero eigenvalue corresponding to mass weighting,

$$P^{\epsilon}_{MM'}(k) = \frac{\delta_D(\ln M - \ln M')}{\bar{n}(M)} - \Theta_{MM'}(k)\frac{MM'}{\bar{\rho}\langle M\rangle_\rho}, \tag{32}$$

where $\langle M\rangle_\rho$ is defined as

$$\langle M\rangle_\rho \equiv \int d\rho(M) M. \tag{33}$$

Here, $\Theta_{MM'}(k)$ is an interpolating function that asymptotes to 1 for $k \to 0$, satisying Eq. (29), while approaching zero in the high-k limit. To be specific, we choose

$$\Theta_{MM'}(k) = [1 + (k[R_L(M) + R_L(M')]/2)^4]^{-1}, \tag{34}$$

where the transition scale is given by the halo Lagrangian radii following our considerations above. Note that, for a covariance of the form Eq. (32), we need $\Theta_{MM'}(k)$ to scale as $1 + \mathcal{O}(k^4)$ in the low-k limit in order to satisfy the conditions Eq. (13) for all k. While different forms of interpolating functions could be chosen, we expect the transition to be related to R_L. The detailed shape of the interpolation is not expected to have a significant impact on the power spectrum prediction, as the transition happens on scales where the power spectrum is still dominated by the deterministic contribution (see Fig. 1).

Equation (32) is certainly not the only possible choice. For example, Ref. [36] derived a covariance given by[3]

$$P^{\epsilon}_{MM'}(k \to 0) = \frac{\delta_D(\ln M - \ln M')}{\bar{n}(M)} - b_1(M)\frac{M'}{\bar{\rho}} - b_1(M')\frac{M}{\bar{\rho}} + b_1(M)b_1(M')\frac{\langle M\rangle_\rho}{\bar{\rho}}. \tag{35}$$

It is easily verified that this ansatz indeed satisfies Eq. (29), assuming the first line in Eq. (13) holds. We could thus multiply the last three terms by the interpolating function $\Theta_{MM'}(k)$ and insert into Eq. (27). However, the additional terms [in particular, the last term in Eq. (35)] grow rapidly towards high k due to their mass weighting, so that they dominate the matter power spectrum for $k \gtrsim 0.5h\,\mathrm{Mpc}^{-1}$ despite the suppression by the interpolating function Eq. (34); this result is insensitive to the shape and steepness of $\Theta_{MM'}(k)$. Thus, we cannot attain our desired high-k limit, which is the standard 1-halo term based on Eq. (31). We instead work with Eq. (32) here, but conclude that simulation measurements of halo stochasticity on large and intermediate scales are essential in order to properly calibrate the halo model prediction.

[3]Note that this was derived using a standard halo model ansatz based on Eq. (31) which does not enforce mass and momentum conservation.

The stochastic contribution to the matter power spectrum [second term in the second line of Eq. (27)] is shown as a green long-dashed line in Fig. 1. We also show the standard 1-halo prediction with its unphysical k^0 behavior at low k (black short-dashed line). As expected, Eq. (32) yields the desired k^4 behavior of the stochastic contribution, while asymptoting to the standard 1-halo term for $k \gtrsim 0.5h\,\mathrm{Mpc}^{-1}$. Qualitatively, this is what one expects for the stochastic contribution in the halo model, although the quantitative behavior for $k \lesssim 1h\,\mathrm{Mpc}^{-1}$ can of course be modified significantly by changing the low-k limit of $P^{\epsilon\epsilon}_{MM'}$ and/or the interpolating function Eq. (34). The result appears roughly consistent with the findings of [23] (e.g. blue curve in Fig. 8 of that reference), who isolated the stochastic contribution to the power spectrum in simulations by subtracting the part correlated with long-wavelength correlations. Interestingly, they find a slightly shallower scaling with k than k^4 even for $k \lesssim 0.1h\,\mathrm{Mpc}^{-1}$. Whether this really implies the existence of another scale much below $k_{\rm NL}$ remains to be seen.

At this point, it is also worth discussing the usual 1-halo vs 2-halo separation. The first term in the second line of Eq. (27) corresponds to the standard 2-halo term. One could refer to the second stochastic term as a 1-halo term, even though it involves a covariance between different halo masses. Alternatively, one could only refer to that part of the stochastic contribution that is proportional to $\delta_D(\ln M - \ln M')$ as a 1-halo contribution, while the remainder of the stochastic part is considered a contribution to a modified 2-halo term (see also [9,41]). In any case, this separation is somewhat arbitrary and a matter of definition, as everything should be derived from the physical assumptions described in Sec. II rather than a separation of the statistics into n-halo terms.

B. Concentration expansion

Let us now turn to the terms in the third and fourth lines of Eq. (27), involving y_c, which come from the perturbative expansion of the concentration c. First, consider the deterministic terms $\propto P_{\rm L}(k)$. Figure 1 shows these terms, assuming $b_1^c = b_1$ which is almost certainly a significant overestimation of the effect, given the fairly small environment dependence observed for the halo concentration in simulations [42]. In fact, this contribution is entirely dominated by the cross term given on the third line of Eq. (27). Clearly, this contribution is significantly smaller and shifted to higher k compared to the terms from the expansion of $\delta_{h,M}$. The main reason for this is that the change of halo profiles due to a change in concentration happens on fairly small scales, of the order of the scale radius of these halos. Further, $|y_c(k,M)|$ is at most $\sim$0.4; that is, halo profiles do not respond strongly to a change in concentration.

Turning to the stochastic terms, we now need a parametrization of $P^{\epsilon c^c}_{MM'}(k)$, scaling as $\mathcal{O}(k^2)$ for $k \to 0$, and $P^{\epsilon^c\epsilon^c}_{MM'}$, which has no low-$k$ constraint. Let us consider the latter. The simplest assumption to make is that each halo's concentration is drawn from a log-normal distribution with fixed scatter $\sigma_{\ln c}$ around the mean relation $\bar{c}(M)$. Then, we have

$$P^{\epsilon^c\epsilon^c}_{MM'} = \frac{\sigma^2_{\ln c}}{\bar{n}}\delta_D(\ln M - \ln M'). \tag{36}$$

The result, using $\sigma_{\ln c} = 0.4$ (of the order of what was found for the scatter in concentration in [42]), is also shown in Fig. 1. Again, we find this to be a small contribution to $P_{mm}(k)$, mainly relevant around $k \sim 2h\,\mathrm{Mpc}^{-1}$. The final remaining term is the stochastic cross-correlation between halo number density and concentration $P^{\epsilon\epsilon^c}_{MM'}(k)$. This is expected to be smaller than the stochastic autocorrelations of halo number and profiles, because it is constrained to vanish on both small and large scales: mass conservation implies a k^4 scaling for $k \to 0$, while for scales $k \gtrsim 1/R_M$ within halos, profiles and number density have to be independent random variables. This means $P^{\epsilon\epsilon^c}_{MM'}(k)$ can only be relevant on a fairly narrow range of scales around $k \sim 1/(R_M + R_{M'})$. For this reason, we do not investigate this term further here.

In summary, in the case of the simple concentration expansion of halo profiles performed here, the effects are suppressed compared to the expansion of $\delta_{h,M}$, so that, depending on the application and range of wave numbers of interest, they can be neglected. We stress however that this assumes that the impact of the large-scale environment on halo profiles is well captured by a change in concentration. If in reality there is a significant effect on the outer regions of halo profiles, then this could make the power spectrum contributions from the profile expansion more significant and push them to larger scales. This is well worth investigating in simulations. We leave this to future work.

IV. MATTER POWER SPECTRUM BEYOND TREE LEVEL

The previous section described the leading order prediction of the halo model, which only matches linear perturbation theory on large scales. Let us now turn to the next higher order incarnation of EHM, where we go to third order in PT. Our goal is to outline the overall features of the result and highlight open issues. We neglect the terms arising from the expansion of the concentration throughout this section.

Let us begin with the deterministic contributions to the matter power spectrum. These can be written as

$$\begin{aligned} P_{mm}(k)|_{\rm det} = &\int d\rho(M)\int d\rho(M')y(k,M)y(k,M') \\ &\times\{b_1(M)b_1(M')[P_{\rm L}(k) + P_{\text{1-loop}}(k)] \\ &+ P^{MM'}_{\rm nlb}(k)\}, \end{aligned} \tag{37}$$

where

$$P_{\text{1-loop}}(k) = \langle \delta^{(2)}(\mathbf{k})\delta^{(2)}(\mathbf{k})\rangle' + 2\langle \delta^{(1)}(\mathbf{k})\delta^{(3)}(\mathbf{k})\rangle' \quad (38)$$

is the 1-loop matter power spectrum [43], and $\delta^{(n)}$ denotes the matter density at nth order in PT. $P^{MM'}_{\text{nlb}}(k)$ contains all nonlinear bias terms that are relevant at 1-loop order,

$$P^{MM'}_{\text{nlb}}(k) = \sum_{\{O,O'\}\neq\{\delta,\delta\}}^{\text{1-loop}} b_O(M) b_{O'}(M') \langle [O](\mathbf{k})[O'](\mathbf{k}')\rangle'. \quad (39)$$

The full expression for the 1-loop halo power spectrum can be found in [44,45]. In analogy with Eq. (38), these terms can be divided into quadratic bias terms which scale similarly to $\langle \delta^{(2)}\delta^{(2)}\rangle$ and cubic bias terms which scale similarly to $\langle \delta^{(1)}\delta^{(3)}\rangle$. The numerically largest term of the former category is given by

$$b_1(M)b_2(M')\langle \delta^{(2)}(\mathbf{k})[\delta^2](\mathbf{k}')\rangle' = b_1(M)b_2(M')\int \frac{d^3\mathbf{q}}{(2\pi)^3} F_2(\mathbf{q},\mathbf{k}-\mathbf{q}) P_{\text{L}}(\mathbf{q}) P_{\text{L}}(\mathbf{k}-\mathbf{q}), \quad (40)$$

where F_2 is the symmetrized second-order perturbation theory kernel [43]. We use the second-order bias derived from the Sheth-Tormen mass function for our results [note that this satisfies Eq. (13)]. At 1-loop order, the third-order renormalized bias contributions to Eq. (39) are all degenerate and can be grouped as a single contribution [44]:

$$\sum_{O=\mathcal{O}(\delta^3)} b_1(M) b_O(M') \langle \delta^{(1)}(\mathbf{k})[O](\mathbf{k})\rangle' = b_1(M) b_{3\text{nl}}(M') \sigma_3^2(k) P_{\text{L}}(k), \quad (41)$$

where $\sigma_3^2(k)$ is a filtered version of the linear power spectrum. The filter is defined explicitly in [44]. For illustrative results, we use the prediction from local Lagrangian biasing [45],

$$b_{3\text{nl}}(M) = \frac{32}{315}[b_1(M) - 1]. \quad (42)$$

Let us also give the expression for the 1-loop propagator, i.e. the cross-correlation of the PT-evolved initial density field and the nonlinear matter density:

$$P_{1m}(k) = \int d\rho(M) y(k,M) \Big\{ b_1(M)[P_{\text{L}}(k) + P_{\text{1-loop}}(k)] + \sum_{O=\mathcal{O}(\delta^2)} b_O(M)\langle \delta^{(2)}(\mathbf{k})[O](\mathbf{k})\rangle' + \sum_{O=\mathcal{O}(\delta^3)} b_O(M)\langle \delta^{(1)}(\mathbf{k})[O](\mathbf{k})\rangle' \Big\}, \quad (43)$$

where the leading contributions to the second and third lines are given by Eqs. (40)–(41) without the factor $b_1(M)$.

Let us finally turn to the stochastic terms at 1 loop, given by

$$P_{mm}(k)|_{\text{stoch}} = \int d\rho(M) \int d\rho(M') y(k,M) y(k,M') \times \left[P^{\epsilon\epsilon}_{MM'}(k) + \int \frac{d^3\mathbf{q}}{(2\pi)^3} P^{\epsilon_\delta\epsilon_\delta}_{MM'}(q) P_{\text{L}}(|\mathbf{k}-\mathbf{q}|) + C_1 + C_2 k^2 \right]. \quad (44)$$

There is only a loop additional contribution to the matter power spectrum which involves $P^{\epsilon_\delta\epsilon_\delta}$ defined following Eq. (28). In order to enforce Eq. (14), we add counterterms C_1 and C_2, whose values are uniquely determined given $P^{\epsilon_\delta\epsilon_\delta}_{MM'}(k)$. These counterterms ensure that the final contribution scales as k^4 in the low-k limit. While evaluating this term requires a parametrization of $P^{\epsilon_\delta\epsilon_\delta}_{MM'}$, we have performed a rough evaluation using a form inspired by the high-k limit discussed in Sec. II D. Including the counterterms, this contribution was found to be of the order of a few percent of the tree-level stochastic term for $k \lesssim 0.5h\,\text{Mpc}^{-1}$. Given the lack of knowledge about $P^{\epsilon_\delta\epsilon_\delta}_{MM'}$ on large and intermediate scales, we do not show it here.

Figure 2 shows the contribution $\propto b_1(M)b_1(M')$ in Eq. (37), as well as the two terms Eq. (40) and Eq. (41) which are a subset of $P^{MM'}_{\text{nlb}}(k)$. Note that, after mass weighting, both Eqs. (40) and (41) yield negative contributions to Eq. (37). We also show the linear and 1-loop matter power spectra as well as the tree-level stochastic term. It is clear that the latter, together with the term scaling as $b_1(M)b_1(M')$, dominate the EHM power spectrum. The nonlinear biases are suppressed by the conservation conditions Eq. (13), so that they only begin to contribute on scales of the order of the halo radius [where $y(k,M)$ begins to be appreciably different from 1]. This suppression is even stronger for terms that scale as $(b_2)^2$.

Thus, although Fig. 2 does not show all EHM contributions, we can already draw some conclusions. For comparison, we also show in Fig. 2 the nonlinear matter power spectrum evaluated for our fiducial cosmology by the Coyote emulator [46], which is accurately calibrated on simulations. This illustrates the well-known fact that $P_{\text{1-loop}}$ overpredicts the true nonlinear power spectrum measured in N-body simulations. The stochastic contribution, at least assuming our parametrization Eq. (32), only exacerbates this problem. The fact that some of the nonlinear bias terms are negative will not solve this issue in the range $k \sim 0.2$–$1h\,\text{Mpc}^{-1}$, as they are too small numerically.

This problem occurs on intermediate scales, which are too small for perturbation theory to be valid, but still larger than the Eulerian radius of halos. Thus, one cannot expect rigorous physical solutions by extrapolating from either

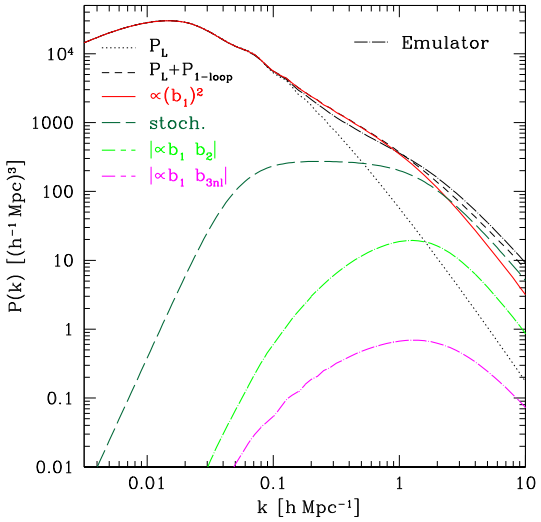

FIG. 2. Illustration of some of the contributions to the 1-loop halo model matter power spectrum Eq. (37) at $z = 0$. The red solid line shows the contribution $\propto b_1(M)b_1(M')$ [terms in brackets in the second line of Eq. (37)]. The green long-dashed line is the same stochastic contribution as in Fig. 1. The light blue dot–long-dashed line shows the contribution from Eq. (40), while the third-order bias contribution [Eq. (41)] is shown as a blue dot-dashed line. The linear power spectrum is shown as a thin dotted line as in Fig. 1, while the thin dashed line shows the matter power spectrum at 1 loop. The thin dot–long-dashed line shows the nonlinear power spectrum from the Coyote emulator [46].

regime. Note that the predictions in the intermediate regime also depend on which perturbative scheme is used, e.g. Eulerian standard perturbation theory (SPT) vs Lagrangian perturbation theory (LPT): different schemes are only guaranteed to give the same result on scales where perturbation theory is valid and will diverge on fully nonlinear scales.

In the framework of the halo model, these issues arise because we assume perturbation theory to describe $\delta_{h,M}$ correctly down to halo scales, which does not hold. For example, nonperturbative effects such as halo exclusion are not included by definition. One approach to address the mismatch on intermediate scales is to perform a matching to simulations. In particular, one can match the halo propagator, which for 1-loop SPT is given in analogy to Eq. (43), by

$$P_{1h}(k,M) = b_1(M)[P_{\rm L}(k) + P_{\text{1-loop}}(k)] + \sum_{O=\mathcal{O}(\delta^2)} b_O(M)\langle\delta^{(2)}(\mathbf{k})[O](\mathbf{k})\rangle' + \sum_{O=\mathcal{O}(\delta^3)} b_O(M)\langle\delta^{(1)}(\mathbf{k})[O](\mathbf{k})\rangle', \quad (45)$$

to simulations by rescaling it with a function $v(k,M)$. Then, by rescaling the profiles $y(k,M) \to y(k,M)v(k,M)$, the matter propagator $P_{1m}(k)$ should be described correctly to the extent that the basic assumption of the halo model is valid. A similar matching of $P_{hh}(k,M,M')$ can be used to determine $P^{\epsilon\epsilon}_{MM'}(k)$.

We leave this to future work, but point out that modifications to the profiles $y(k,M)$ as a way to fit intermediate scales have also been proposed in [27], who expanded the profile on large scales in powers of k^2. Reference [23] discussed a subtraction of the high-k contribution of the loop integrals to Eq. (38) as an effective profile.

Common to all these attempts at solving the issue of intermediate scales is the fact that we need to introduce another scale that is larger than the typical halo size (and close to the nonlinear scale). This is clear from Fig. 2, and given that the typical wave number corresponding to halo radii is of the order of $k_{\rm HM} \sim \pi/R_{\rm HM} \sim 8h\,{\rm Mpc}^{-1}$. The necessity of adding nonperturbative terms involving a new scale not directly related to halo profiles breaks with the philosophy of the halo model as outlined in the Introduction, i.e. that predictions should be given completely in terms of perturbation theory and well-defined properties of halos. Moreover, the distinction between deterministic and stochastic contributions is blurred in this transition regime. Nevertheless, the goal is to sufficiently constrain the additional terms to keep the halo model predictive, in particular, by considering various statistics such as matter and halo power spectra and bispectra. We leave the whole question of intermediate scales as an open issue for future work.

V. BISPECTRUM

We now consider the bispectrum (three-point function) of matter, and present the EHM prediction at tree level. Given our findings from Sec. III, we drop the terms coming from the concentration expansion, simplifying the expressions considerably. At tree level, we then need to expand Eq. (3) to second order. This yields

$$\delta_{h,M}(\mathbf{x},\tau) = b_1(M,\tau)\delta(\mathbf{x},\tau) + \frac{1}{2}b_2(M,\tau)[\delta^2](\mathbf{x},\tau) + \frac{1}{2}b_{s^2}(M,\tau)[(s_{ij})^2](\mathbf{x},\tau) + [\epsilon](M,\mathbf{x},\tau) + [\epsilon_\delta\delta](M,\mathbf{x},\tau), \quad (46)$$

where we have slightly changed notation to match standard convention for the density biases, and $s_{ij} \equiv (3\Omega_m\mathcal{H}^2/2)^{-1}(\partial_i\partial_j - \delta_{ij}\nabla^2)\Phi$ is a scaled version of the tidal field. The resulting expression for the matter bispectrum is given in Eq. (C1).

The two main differences to commonly used halo model bispectra are that first, we are including the two second-order bias terms, with respect to density squared and tidal field squared [third and fourth lines in Eq. (C1)]. The tidal bias [47,48] has not been included in halo model calculations of the bispectrum so far (e.g. [19,49]), but is

straightforward to include once a parametrization of $b_{s^2}(M)$ is given which satisfies Eq. (13).

Second, the general stochastic terms which need to be included at this order are given in the last line of Eq. (C1). Let us repeat them here:

$$B_{mmm}(k_1,k_2,k_3) \overset{\text{stoch}}{=} \int d\rho(M_1)\int d\rho(M_2)\int d\rho(M_3) \times y(k_1,M_1)y(k_2,M_2)y(k_3,M_3) \times \{B^{\epsilon}_{M_1M_2M_3}(k_1,k_2,k_3) + [P_{\rm L}(k_2)P^{\epsilon_\delta\epsilon}_{M_1M_3}(k_3) + P_{\rm L}(k_3)P^{\epsilon_\delta\epsilon}_{M_1M_2}(k_2)] + 2\,\text{perm}\}, \quad (47)$$

where

$$B^{\epsilon}_{M_1M_2M_3}(k_1,k_2,k_3) \equiv \langle [\epsilon](M_1,k_1)[\epsilon](M_2,k_2)[\epsilon](M_3,k_3)\rangle' \qquad P^{\epsilon_\delta\epsilon}_{M_1M_2}(k) \equiv \langle [\epsilon_\delta](M_1,\mathbf{k})[\epsilon](M_2,\mathbf{k}')\rangle'. \quad (48)$$

We thus need a parametrization of $B^{\epsilon\epsilon\epsilon}_{M_1M_2M_3}$ and $P^{\epsilon_\delta\epsilon}_{M_1M_2}$ (since we are working at tree level, we do not need to perform any renormalization on ϵ and $\epsilon_\delta\delta$). In the high-k limit, we can use the prediction of Poisson sampling from the local deterministic halo abundance, Eq. (21). At low k, Eq. (13) requires that the mass-weighted integral of these quantities over any of the masses M_i vanishes. We can immediately generalize our interpolating ansatz Eq. (32) to $P^{\epsilon_\delta\epsilon}_{M_1M_2}$ through

$$P^{\epsilon_\delta\epsilon}_{M_1M_2}(k) = \frac{1}{2}\frac{b_1(M_1)}{\bar{n}(M_1)}\delta_D(\ln M_1 - \ln M_2) - \frac{1}{2}\Theta_{M_1M_2}(k)\frac{M_1M_2}{\bar{\rho}\langle b_1 M\rangle_\rho}b_1(M_1)b_1(M_2), \quad (49)$$

where

$$\langle b_1 M\rangle_\rho \equiv \int d\rho(M) b_1(M) M. \quad (50)$$

In Appendix C we also give a somewhat more lengthy expression for $B^{\epsilon\epsilon\epsilon}$ which satisfies the corresponding constraints [Eq. (C4)]. While it is a simple extension of Eq. (32), it clearly is not the only possible choice. Again, we stress that further numerical investigations of halo stochasticity, including its three-point function, as a function of scale are necessary in order to obtain an accurate halo model bispectrum.

Nevertheless, via Eqs. (49) and (C4), and given bias parameters $b_2(M)$, $b_{s^2}(M)$, Eq. (C1) yields a consistent matter bispectrum obeying all symmetries of the matter density, and asymptoting to the tree-level matter bispectrum on large scales. It does not include the effect of a modulation of halo profiles by large-scale density perturbations, as we have found it to be numerically small in the case of the power spectrum, but this can be easily added back in. Of course, we expect the same issues on intermediate scales to arise that appear for the power spectrum.

VI. MATTER VELOCITY FIELD

So far, we have only considered the matter density field, since this is phenomenologically the most important quantity for large-scale structure. Let us now consider how the nonlinear matter velocities are described in the self-consistent halo model approach pursued here. First of all, since the single-stream fluid picture breaks down on nonlinear scales, our goal has to be to derive the velocity *distribution* at a given point $(\mathbf{x},\tau)$.

Let us denote the matter velocity predicted by perturbation theory, at the relevant order used in the halo model, by $\mathbf{v}_{\rm PT}$. As argued in Ref. [29], halo velocities are unbiased with respect to matter velocities up to higher derivative terms; that is, the velocity of the effective halo fluid obtained by coarse graining the halo distribution is given by

$$\mathbf{v}_{h,M}(\mathbf{x},\tau) = \mathbf{v}_{\rm PT}(\mathbf{x},\tau) + \mathcal{O}(\nabla^2\mathbf{v}_{\rm PT}, \boldsymbol{\nabla}\delta_{\rm PT}). \quad (51)$$

Since these higher derivative terms are assumed to be given entirely by the halo profiles in our approach, we set $\mathbf{v}_{h,M} = \mathbf{v}_{\rm PT}$. Then, the velocity distribution at $(\mathbf{x},\tau)$ in the halo model is given by[4]

$$P(\mathbf{v};\mathbf{x}) = \int dn(M)\int d^3\mathbf{y}[1+\delta_{h,M}(\mathbf{y})] \times P_{v,h}(\mathbf{v}-\mathbf{v}_{\rm PT}(\mathbf{x});M,\mathbf{x}-\mathbf{y}), \quad (52)$$

where we have dropped the time argument for clarity, and $\int dn(M) \equiv (\int \bar{n}(M) d\ln M)^{-1}\int \bar{n} d\ln M$ is the normalized integral over the halo number density (that is, without mass weighting). $P_{v,h}(\mathbf{v};M,\mathbf{r})$ denotes the mean normalized velocity distribution within halos of mass M at radius $\mathbf{r}$. By construction, this obeys

$$\int d^3\mathbf{v} P_{v,h}(\mathbf{v};M,\mathbf{r}) = 1 \quad \text{and} \quad \int d^3\mathbf{v}\mathbf{v} P_{v,h}(\mathbf{v};M,\mathbf{r}) = 0. \quad (53)$$

For spherically symmetric halos, $P_{v,h}$ can be written as

$$P_{v,h}(\mathbf{v};M,\mathbf{r}) = P_v(v_\parallel = \mathbf{v}\cdot\hat{\mathbf{r}}; v_\perp = |\mathbf{v} - (\mathbf{v}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}}|; M, r),$$

i.e. in terms of the joint distribution of radial and tangential velocities. See [50,51] for examples of modeling this velocity distribution.

[4]Here, we are working in the nonrelativistic limit and ignore corrections of the order of $\mathbf{v}^2$.

Equation (52) can then be generalized by allowing for a dependence of the velocity distribution on the halo concentration, for example, leading to an expansion analogous to that discussed in Sec. II B. Further, one can straightforwardly apply the same type of reasoning to obtain the momentum density, or mass-weighted velocity.

VII. CONNECTION TO THE EFT

The EHM approach described in Sec. II predicts, by construction, a matter density field which matches the results of perturbation theory to any desired order on large scales. Beyond the large-scale limit, the halo profiles lead to higher derivative terms $\propto \nabla^2\delta$, $(\nabla\delta)^2$, and so on. Further, the halo model contains a stochastic contribution to the matter density field, i.e. a contribution which does not correlate with long-wavelength perturbations. All these contributions satisfy the requirement of large-scale mass and momentum conservation as long as the conditions Eq. (13) are satisfied.

In this sense, this halo model approach consistently extends the predictions of EFT of LSS to nonlinear scales, which necessarily implies that the halo model is not guaranteed to be within a well-defined theoretical uncertainty from the true answer when going beyond perturbative scales $k/k_{\rm NL} \ll 1$. A detailed study of the connections of the halo model to the EFT, while interesting, is beyond the scope of this paper. We just make two comments of general interest here.

Matching to EFT parameters: As emphasized in Sec. II, one key virtue of EHM is that it can be taken beyond tree level. In the form that we have defined the implementation there, corrections to the perfect fluid description, i.e. the terms added by the EFT, are exclusively provided by the halo profiles. The results in Sec. IV already show however that the EHM ansatz fails to even roughly predict parameters such as the effective sound speed c_s: EHM predicts a scale $k_{\rm HM} = \pi/R_{\rm HM}$, while simulation measurements (e.g. [52]) find that the correct scale is $k_{\rm NL} \ll k_{\rm HM}$. There is no reason to expect that this problem will be solved by higher loops; instead one has to separately model the transition regime as discussed at the end of Sec. IV.

Higher derivative terms: In the halo model, δ_m is written as a convolution of a scalar $\delta_{h,M}$ with a profile $y(r|M)$, where the profile is assumed to generate all higher derivative terms. For this reason, we only obtain higher derivative terms of the type $\partial^2 O$ and $\partial_i O \partial^i O'$, where O, O' are scalar operators appearing in the expansions Eqs. (3) and (7). The second type is generated by having both $\partial^2(OO')$ and $O(\partial^2 O')$ in the expansion. This also holds when including the dependence of halo profiles and triaxiality on long-wavelength perturbations. Hence, the halo model, as described in Sec. II, does not generate higher derivative terms of the form $\partial_i s_{jk} \partial^k s^{ij}$ or similar terms for other nonscalar operators, which are, in general, present in the EFT. This is a prediction which can be tested on simulations, by comparing the measured amplitude (on scales within the perturbative regime) of higher derivative terms of the type $\partial_i s_{jk} \partial^k s^{ij}$ with, for example, $\partial_i s_{jk} \partial^i s^{jk}$. The halo model as described here only produces the second term. Of course, it is always possible to explicitly include any higher derivative term in the expansion of the halo overdensity Eq. (3).

VIII. CONCLUSIONS

We have presented a general procedure (EHM) for constructing a halo model description of the nonlinear large-scale structure which guarantees mass and momentum conservation on large scales. This procedure allows for perturbation theory results to be matched to any given order. Finally, a single set of input ingredients (mass function, bias parameters, profiles, and stochasticity covariances) describes all matter and halo auto and cross correlations.

We have attempted to write down the most general expression for the matter density field that follows from the basic halo model assumption stated at the beginning of Sec. I and that remains predictive. For this reason, we have only allowed terms at lowest order in derivatives in the halo density expansion Eq. (3), thereby declaring that halo profiles are responsible for all higher derivative terms in the matter density field. While the number of input parameters in the model increases as one goes to higher order (in particular, the bias parameters of halos), these parameters can be measured in simulations (e.g. [44,45,47,53]) or predicted via the peak-background split approach for example. The key virtues of the halo model, namely simple, numerically cheap predictions for nonlinear matter statistics on all scales that are physically motivated, are retained in any case.

The new ingredients discussed here for the first time are the halo stochasticity covariance, the concentration expansion allowing for the dependence of halo profiles on the environment, and a clarification of the impact of low-mass halos on halo model predictions. The last point, discussed in detail in Appendix A, also applies to existing formulations of the halo model.

Perhaps the most important conclusion of this work is that the halo stochasticity covariance is a key ingredient of the halo model, and likely to be numerically important in the transition region between the classic 2-halo and 1-halo regimes. This quantity has clearly not been studied in sufficient detail so far, with the most detailed studies being Refs. [36,38]. Here, we have described a general procedure to derive the high-k limit of the stochasticity (in the 1-halo regime) in terms of the perturbative bias parameters and mass function.

The prescription given here does not by itself address the failure of the halo model to describe the transition region between PT scales and the 1-halo regime. In fact, going to 1-loop order in the power spectrum, we found that the halo

model performs worse than perturbation theory on scales $k \sim 0.2$–$1h\,\mathrm{Mpc}^{-1}$. This will most likely require additional ingredients (see e.g. [23,27] for related approaches). We leave this as a major open problem for future work.

Turning to halo profiles, we have allowed for the spherically averaged halo profiles as well as halo triaxiality to depend on long-wavelength perturbations. In order to avoid many free functions of scale, we have parametrized this dependence only through the concentration. This however can easily be augmented to include the environmental dependence of halo outskirts as well. Interestingly, we found that halo triaxiality is likely to be unimportant in practice, as it is largely degenerate with the expansion of the spherically averaged profiles.

A detailed comparison of the halo model power spectrum and bispectrum with simulation results is left for future work. This will also necessitate more study of the halo stochasticity. In order to be a fair comparison, this has to make use of state-of-the-art numerically calibrated halo mass function, biases, and profiles.

Let us also briefly discuss assembly bias, i.e. the fact that the large-scale properties of halos depend on more than just the halo mass (e.g. [54–56]). In principle, assembly bias can be straightforwardly included in the halo model, by promoting the integral over mass in Eq. (12) to a multidimensional integral over mass, formation time, and/or other quantities. Correspondingly, the mass function $\bar{n}$, mean concentration $\bar{c}$, bias parameters b_O and b_O^c, as well as stochastic fields ϵ_O, ϵ_O^c all become functions of mass, formation time, and so on. Note that assembly bias can only affect halo model predictions if both profiles and biases and/or stochastic fields depend on additional variables, for example, if at fixed mass halos with higher concentration are more biased. These effects thus only become relevant in the intermediate to 1-halo regime.

Finally, the halo model can also be generalized to a model for galaxy statistics via the halo occupation distribution (HOD) approach. In the spirit of the approach described here, the HOD for halos of a given mass should also be allowed to depend on the long-wavelength perturbations via an expansion of the same type as in Eq. (3). Of course, assembly bias effects can also be included as described just above. These are expected to be more important for galaxy clustering than for the matter density field; for example, certain types of galaxies may live preferentially in early- or late-forming halos. We leave this to future work.

ACKNOWLEDGMENTS

I would like to thank Tobias Baldauf, Mehrdad Mirbabayi, Emmanuel Schaan, Uroš Seljak, Masahiro Takada, and Matias Zaldarriaga for many helpful comments and discussions. I gratefully acknowledge support from the Marie Curie Career Integration Grant No. FP7-PEOPLE-2013-CIG "FundPhysicsAndLSS."

APPENDIX A: ON THE LOW-MASS CUTOFF OF HALOS

The halo model is based on parametrizations of the abundance, bias parameters, and profiles of halos, all as a function of mass. Clearly, these are only calibrated over a certain mass range in simulations. At high masses, there is no obstacle, in principle, to measuring halo properties accurately. A practical issue is that halos become exponentially rare at very high masses. However, this also makes them phenomenologically unimportant. For this reason, any extrapolation used at high masses is likely to be well under control.

On the other hand, properties of low-mass halos are poorly constrained by simulations due to resolution limits. The mass-weighting integrals, for example in Eq. (13), converge very slowly towards low masses. This raises the question of whether the halo model predictions actually rely on extremely low-mass halos whose properties are poorly known.

Here, we show that this is not the case. Consider the case where the mass function, bias and profiles are well calibrated to a minimum mass M_s. We show that the uncertainties to the halo model predictions introduced by halos of mass below M_s are of the order of $(kR_{M_s})^2$. If the scales of interest are $k \ll 1/R_{M_s}$, then this is a negligible uncertainty on the halo model predictions.

To prove this, we introduce a low-mass cutoff M_s so that all mass-weighting integrals become

$$\int d\rho(M) \to \int_{\ln M_s}^{\infty} d\ln M \frac{M}{\bar{\rho}} \bar{n}(\ln M). \tag{A1}$$

In order to fix global mass conservation, we add an effective term to the mass function at the cutoff,

$$\bar{n}(M) \to \bar{n}(M) + \bar{n}_s \delta_D(\ln M - \ln M_s), \tag{A2}$$

where $\bar{n}_s$ is determined by requiring

$$\int_{\ln M_s} d\ln M \frac{M}{\bar{\rho}} \bar{n}(\ln M) + \frac{M_s}{\bar{\rho}} \bar{n}_s = 1. \tag{A3}$$

Similarly, in order to ensure the consistency condition for b_1 [Eq. (13)], we let

$$b_1(M) = \begin{cases} b_1(M) & M > M_s \\ b_{1s} & M = M_s, \end{cases} \tag{A4}$$

and require

$$\int_{\ln M_s} d\ln M \frac{M}{\bar{\rho}} \bar{n}(M) b_1(M) + \frac{M_s}{\bar{\rho}} \bar{n}_s b_{1s} = 1. \tag{A5}$$

Corresponding conditions are to be placed on the other biases $b_O(M)$. For simplicity, we restrict ourselves to the

matter statistics in the linear version of EHM here. The deterministic contributions to $P_{1m}(k)$ and $P_{mm}(k)$ involve the following integral:

$$\begin{aligned}&\int d\rho(M) b_1(M) y(k,M)\\ &\to \int_{\ln M_s} d\ln M \frac{M}{\bar\rho}\bar n(M) b_1(M) y(k,M) + \frac{M_s}{\bar\rho}\bar n_s b_{1s} y(k,M_s)\\ &\overset{kR_{M_s}\ll 1}{=} 1 - k^2\left[\int_{\ln M_s} d\ln M \frac{M}{\bar\rho}\bar n(M) b_1(M) a_M R_M^2\right.\\ &\left. + \frac{M_s}{\bar\rho}\bar n_s b_{1s} a_{M_s} R_{M_s}^2\right],\end{aligned} \tag{A6}$$

where in the last line we have used Eq. (A4) and the low-k limit of the profile Eq. (9). Taking the derivative with respect to $\ln M_s$ of this expression, it is easy to verify that this result is independent of M_s up to corrections of the order of $(kR_{M_s})^2$.

We now consider the stochastic contribution. In order to satisfy Eq. (13), we add an additional stochastic field ϵ_s which only contributes to the stochasticity of halos of mass M_s. We require $\epsilon_s(\mathbf{k})$ to satisfy, in the same sense as Eq. (13),

$$\int_{\ln M_s} d\ln M \frac{M}{\bar\rho}\bar n(M)\epsilon(M,\mathbf{k}) + \frac{M_s}{\bar\rho}\bar n_s \epsilon_s(\mathbf{k}) = \mathcal{O}(R_{M_s}^2 k^2). \tag{A7}$$

Since ϵ_s is supposed to describe halos of mass $\le M_s$, we require the scaling in terms of R_{M_s} given on the rhs of Eq. (A7). Taking the autocorrelation of this equation then implies

$$\begin{aligned}&\left\langle\left[\int_{\ln M_s} d\ln M \frac{M}{\bar\rho}\bar n(M)\epsilon(M,\mathbf{k}) + \frac{M_s}{\bar\rho}\bar n_s\epsilon_s(\mathbf{k})\right]\right.\\ &\left.\times\left[\int_{\ln M_s} d\ln M' \frac{M'}{\bar\rho}\bar n(M')\epsilon(M',\mathbf{k}') + \frac{M_s}{\bar\rho}\bar n_s\epsilon_s(\mathbf{k}')\right]\right\rangle'\\ &\propto (R_{M_s}k)^4.\end{aligned} \tag{A8}$$

The stochastic contribution to $P_{mm}(k)$,

$$\int d\rho(M)\int d\rho(M') P_\epsilon^{MM'}(k) y(k,M) y(k,M'), \tag{A9}$$

can then easily be shown, via Eqs. (9) and (A7), to scale as k^4 and depend on M_s only through terms of the order of $(R_{M_s}k)^2$.

We conclude that, for $k < 1/R_{M_s}$ where M_s is the lowest mass for which halo properties are well calibrated, the halo model predictions are under accurate theoretical control.

APPENDIX B: HALO TRIAXIALITY

Dark matter halos are triaxial, and the orientation of the axes, as well as having a random component, correlates with large-scale tidal fields s_{ij}. At linear order in the tidal field, this coupling can generally be of the form, dictated by symmetry,

$$y(\mathbf{r},M,\tau,c)|_{s_{ij}} = \left[1 + f_s^c(r,M,\tau) s_{ij}\frac{r^i r^j}{r^2}\right] y(r,M,\tau,c), \tag{B1}$$

where $f_s^c(r,M,\tau)$ is a general function. Of course, in order to retain the predictivity of the halo model, we would like to reduce this to a number in analogy to the concentration expansion introduced above. One possible choice would be to assume that the tidal field distorts halos in a homologous way,

$$y(\mathbf{r},M,\tau,c)|_{s_{ij}} = y\left(\sqrt{r^2 + b_s^c(M,\tau) s_{ij} r^i r^j}, M,\tau,c\right), \tag{B2}$$

which implies $f_s^c = b_s^c(\partial\ln y/\partial\ln r)/2$. However, the integral over this quantity does not vanish, and thus violates the constraint Eq. (5). Let us thus instead choose, for illustrative purposes,

$$\begin{aligned}y(\mathbf{r},M,\tau,c)|_{s_{ij}} &= y(r,M,\tau,c)\\ &+ b_s^c(M,\tau) s_{ij}\frac{\partial^i\partial^j}{\partial^2} y_c(r,M,\tau,c),\end{aligned} \tag{B3}$$

which satisfies Eq. (5). Note that for typical universal halo profiles (such as NFW or Einasto) the functions y_c and $\partial y/\partial\ln r$ are very similar. Equation (B3) is sufficient at linear order, but can be extended to higher order in the same way as Eq. (6), including all terms that have the same trace-free symmetric structure. At quadratic order, this will involve δs_{ij}, $s_i^k s_{kj} - \delta_{ij}(s_{kl})^2/3$ and $\epsilon_s^t s_{ij}$.

Now, when written in the form Eq. (B3), the terms from b_s^c and higher order tidal coupling are all degenerate with terms in the concentration bias expansion. This is because higher derivatives are always contracted with the s_{ij}, i.e. $\partial_i\partial_j s^{ij}$, $\partial_i s^{ik}\partial^j s_{jk}$ and so on, which brings them into the form $\partial^2\delta$, $(\partial_i\delta)^2$, etc. However, what if we allow the tidal coupling to have a different r dependence than the specific form $(\partial_i\partial_j/\partial^2)y_c$? At low k, the Fourier space version of $\partial y/\partial s_{ij}$, when enforcing mass conservation, has to be given by

$$\frac{\partial y}{\partial s_{ij}}(k,M) \overset{k\to 0}{=} a R_M^2 k_i k_j + \mathcal{O}(k^4), \tag{B4}$$

where a is a constant. Again, this will lead to the same term at leading order as the concentration expansion.

Finally, we should also take into account random triaxiality of halos; this was the case studied by [34]. This can be achieved simply by replacing s_{ij} in the relations above with a stochastic trace-free tensor field ϵ^t_{ij}. The conclusions remain the same: these terms scale in a very similar way as the stochastic terms in the concentration expansion.

Thus, only by choosing a functional form for $\partial y/\partial s_{ij}$ that is significantly different from the profile expansion $\partial y/\partial \ln c$ can the tidal coupling of halo triaxiality produce a significant difference to the terms already included in the concentration expansion. Moreover, this difference will only appear at fairly high k. It thus seems likely that halo triaxiality will be a subdominant component of the halo model. The terms found by [34] in the halo model matter bispectrum would thus be effectively captured, in our formulation, by second and third moments of ϵ^c and ϵ^c_δ (as well as their cross-correlations with ϵ_δ and ϵ, respectively). Note that we have not written these terms in Sec. V.

APPENDIX C: HALO MODEL BISPECTRUM

Using the second-order bias expansion Eq. (46), and neglecting the concentration expansion, we obtain the following result for the matter bispectrum in the halo model:

$$\begin{aligned} B_{mmm}(k_1,k_2,k_3) = &\int d\rho(M_1)\int d\rho(M_2)\int d\rho(M_3) y(k_1,M_1)y(k_2,M_2)y(k_3,M_3)\\ &\times\Big\{ b_1(M_1)b_1(M_2)b_1(M_3)B_{\rm T}(k_1,k_2,k_3) + b_2(M_1)b_1(M_2)b_1(M_3)P_{\rm L}(k_2)P_{\rm L}(k_3) + 2\,\text{perm}\\ &+ b_{s^2}(M_1)b_1(M_2)b_1(M_3)\left[(\hat{\mathbf{k}}_2\cdot\hat{\mathbf{k}}_3)^2-\frac{1}{3}\right]P_{\rm L}(k_2)P_{\rm L}(k_3) + 2\,\text{perm}\\ &+ B^{\epsilon}_{M_1M_2M_3}(k_1,k_2,k_3) + [P_{\rm L}(k_2)P^{\epsilon_\delta\epsilon}_{M_1M_3}(k_3) + P_{\rm L}(k_3)P^{\epsilon_\delta\epsilon}_{M_1M_2}(k_2)] + 2\,\text{perm}\Big\}, \end{aligned} \tag{C1}$$

where the tree-level matter bispectrum is given by

$$B_{\rm T}(k_1,k_2,k_3) = 2F_2(\mathbf{k}_1,\mathbf{k}_2)P_{\rm L}(k_1)P_{\rm L}(k_2) + 2\,\text{perm}, \tag{C2}$$

and the stochastic terms are defined in Eq. (48).

One possible form of $B^{\epsilon\epsilon\epsilon}$ that satisfies Eq. (13) in the low-k limit, i.e.

$$\int d\rho(M_1) B^{\epsilon\epsilon\epsilon}_{M_1M_2M_3}(k_1,k_2,k_3) \overset{k_1\to 0}{=} 0, \tag{C3}$$

can be constructed as follows:

$$\begin{aligned} B^{\epsilon}_{M_1M_2M_3}(k_1,k_2,k_3) = &\frac{\delta_D(\ln M_1-\ln M_2)\delta_D(\ln M_1-\ln M_3)}{[\bar{n}(M_1)]^2}\\ &+\Theta_{M_1M_2M_3}(k)\left[-\frac{M_1M_2}{\bar{\rho}\langle M\rangle_\rho}\frac{\delta_D(\ln M_2-\ln M_3)}{\bar{n}(M_2)} + \frac{M_1}{\bar{\rho}^2\langle M\rangle^2_\rho}\left(M_1-\frac{1}{3}\frac{\langle M^2\rangle_\rho}{\langle M\rangle_\rho}\right)M_2M_3\right.\\ &\left.+\,2\,\text{cyclic perm}\right]. \end{aligned} \tag{C4}$$

Here, we have defined $\langle M^2\rangle_\rho \equiv \int d\rho(M)M^2$, and generalized Eq. (34) to

$$\Theta_{M_1M_2M_3}(k) = [1+(k[R_{M_1}+R_{M_2}+R_{M_3}])^4]^{-1}. \tag{C5}$$

[1] A. Cooray and R. K. Sheth, Phys. Rep. **372**, 1 (2002).
[2] M. Takada and B. Jain, Mon. Not. R. Astron. Soc. **344**, 857 (2003).
[3] M. Takada and T. Hamana, Mon. Not. R. Astron. Soc. **346**, 949 (2003).
[4] W. Hu and B. Jain, Phys. Rev. D **70**, 043009 (2004).
[5] A. R. Zentner, D. H. Rudd, and W. Hu, Phys. Rev. D **77**, 043507 (2008).
[6] F. Schmidt, M. Lima, H. Oyaizu, and W. Hu, Phys. Rev. D **79**, 083518 (2009).
[7] M. Takada and B. Jain, Mon. Not. R. Astron. Soc. **395**, 2065 (2009).
[8] F. Schmidt, W. Hu, and M. Lima, Phys. Rev. D **81**, 063005 (2010).
[9] R. E. Smith, V. Desjacques, and L. Marian, Phys. Rev. D **83**, 043526 (2011).
[10] I. Kayo, M. Takada, and B. Jain, Mon. Not. R. Astron. Soc. **429**, 344 (2013).
[11] M. Takada and W. Hu, Phys. Rev. D **87**, 123504 (2013).
[12] L. Lombriser, K. Koyama, and B. Li, J. Cosmol. Astropart. Phys. 03 (2014) 021.
[13] A. Barreira, B. Li, W. A. Hellwing, L. Lombriser, C. M. Baugh, and S. Pascoli, J. Cosmol. Astropart. Phys. 04 (2014) 029.
[14] Y. Li, W. Hu, and M. Takada, Phys. Rev. D **89**, 083519 (2014).
[15] E. Massara, F. Villaescusa-Navarro, and M. Viel, J. Cosmol. Astropart. Phys. 12 (2014) 053.
[16] M. P. van Daalen and J. Schaye, Mon. Not. R. Astron. Soc. **452**, 2247 (2015).
[17] J. Liu and J. C. Hill, Phys. Rev. D **92**, 063517 (2015).
[18] A. Kuntz, Astron. Astrophys. **584**, A53 (2015).
[19] A. Lazanu, T. Giannantonio, M. Schmittfull, and E. P. S. Shellard, arXiv:1510.04075.
[20] A. J. Mead, J. A. Peacock, C. Heymans, S. Joudaki, and A. F. Heavens, Mon. Not. R. Astron. Soc. **454**, 1958 (2015).
[21] A. Paranjape, K. Kovac, W. G. Hartley, and I. Pahwa, Mon. Not. R. Astron. Soc. **454**, 3030 (2015).
[22] S. Tassev and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 12 (2012) 011.
[23] T. Baldauf, E. Schaan, and M. Zaldarriaga, arXiv:1507.02255.
[24] D. Baumann, A. Nicolis, L. Senatore, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 07 (2012) 051.
[25] P. Valageas and T. Nishimichi, Astron. Astrophys. **527**, A87 (2011).
[26] P. Valageas and T. Nishimichi, Astron. Astrophys. **532**, A4 (2011).
[27] I. Mohammed and U. Seljak, Mon. Not. R. Astron. Soc. **445**, 3382 (2014).
[28] U. Seljak and Z. Vlah, Phys. Rev. D **91**, 123516 (2015).
[29] M. Mirbabayi, F. Schmidt, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 07 (2015) 030.
[30] R. Angulo, M. Fasiello, L. Senatore, and Z. Vlah, J. Cosmol. Astropart. Phys. 09 (2015) 029.
[31] P. McDonald, Phys. Rev. D **74**, 103512 (2006).
[32] V. Assassi, D. Baumann, D. Green, and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 08 (2014) 056.
[33] J. F. Navarro, C. S. Frenk, and S. D. M. White, Astrophys. J. **490**, 493 (1997).
[34] R. E. Smith, P. I. R. Watts, and R. K. Sheth, Mon. Not. R. Astron. Soc. **365**, 214 (2006).
[35] U. Seljak, N. Hamaus, and V. Desjacques, Phys. Rev. Lett. **103**, 091303 (2009).
[36] N. Hamaus, U. Seljak, V. Desjacques, R. E. Smith, and T. Baldauf, Phys. Rev. D **82**, 043515 (2010).
[37] N. Hamaus, U. Seljak, and V. Desjacques, Phys. Rev. D **86**, 103513 (2012).
[38] Y.-C. Cai, G. Bernstein, and R. K. Sheth, Mon. Not. R. Astron. Soc. **412**, 995 (2011).
[39] R. K. Sheth and G. Tormen, Mon. Not. R. Astron. Soc. **308**, 119 (1999).
[40] J. S. Bullock, T. S. Kolatt, Y. Sigad, R. S. Somerville, A. V. Kravtsov, A. A. Klypin, J. R. Primack, and A. Dekel, Mon. Not. R. Astron. Soc. **321**, 559 (2001).
[41] T. Baldauf, U. Seljak, R. E. Smith, N. Hamaus, and V. Desjacques, Phys. Rev. D **88**, 083507 (2013).
[42] A. V. Macciò, A. A. Dutton, F. C. van den Bosch, B. Moore, D. Potter, and J. Stadel, Mon. Not. R. Astron. Soc. **378**, 55 (2007).
[43] F. Bernardeau, S. Colombi, E. Gaztañaga, and R. Scoccimarro, Phys. Rep. **367**, 1 (2002).
[44] P. McDonald and A. Roy, J. Cosmol. Astropart. Phys. 08 (2009) 020.
[45] S. Saito, T. Baldauf, Z. Vlah, U. Seljak, T. Okumura, and P. McDonald, Phys. Rev. D **90**, 123522 (2014).
[46] K. Heitmann, E. Lawrence, J. Kwan, S. Habib, and D. Higdon, Astrophys. J. **780**, 111 (2014).
[47] K. C. Chan, R. Scoccimarro, and R. K. Sheth, Phys. Rev. D **85**, 083509 (2012).
[48] T. Baldauf, U. Seljak, V. Desjacques, and P. McDonald, Phys. Rev. D **86**, 083540 (2012).
[49] R. E. Smith, R. Scoccimarro, and R. K. Sheth, Phys. Rev. D **75**, 063512 (2007).
[50] R. E. Smith, R. K. Sheth, and R. Scoccimarro, Phys. Rev. D **78**, 023523 (2008).
[51] T. Y. Lam, F. Schmidt, T. Nishimichi, and M. Takada, Phys. Rev. D **88**, 023012 (2013).
[52] J. J. M. Carrasco, S. Foreman, D. Green, and L. Senatore, J. Cosmol. Astropart. Phys. 07 (2014) 057.
[53] T. Lazeyras, C. Wagner, T. Baldauf, and F. Schmidt, J. Cosmol. Astropart. Phys. 02 (2014) 018.
[54] L. Gao, V. Springel, and S. D. M. White, Mon. Not. R. Astron. Soc. **363**, L66 (2005).
[55] R. H. Wechsler, A. R. Zentner, J. S. Bullock, A. V. Kravtsov, and B. Allgood, Astrophys. J. **652**, 71 (2006).
[56] N. Dalal, M. White, J. R. Bond, and A. Shirokov, Astrophys. J. **687**, 12 (2008).