Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **502,** 5495–5507 (2021)
Advance Access publication 2021 February 15

doi:10.1093/mnras/stab445

# Counts-in-cells of subhaloes in the IllustrisTNG simulations: the role of baryonic physics

Christine C. Dantas★

*Divisão de Astrofísica (INPE-MCTI), 12227-010 São José dos Campos, SP, Brazil*

Accepted 2021 February 11. Received 2021 February 11; in original form 2020 July 13

## ABSTRACT

We present an analysis of the counts-in-cells (CiC) statistics of subhaloes in the publicly available IllustrisTNG cosmological simulations (TNG100-1, TNG100-3, and TNG300-3), considering their full and dark-only versions, in redshifts ranging from $z = 0$ to $z = 5$, and different cell sizes. We evaluated two CiC models: the gravitational quasi-equilibrium distribution (GQED) and the negative binomial distribution (NBD), both presenting good fits, with small detectable differences in the presence of baryons. Scaling and time dependencies of the best-fitting parameters showed similar trends compared with the literature. We derived a matter density-in-cells probability distribution function (PDF), associated with the GQED, which was compared to the PDF proposed by Uhlemann et al., for the IllustrisTNG 100-3-Dark run at $z = 0$. Our results indicate that the simplest gravithermodynamical assumptions of the GQED model hold in the presence of baryonic dissipation. Interestingly, the smoothed (density-in-cells) version of the GQED is also adequate for describing the dark matter one-point statistics of subhaloes and converges, to subpercentage levels (for an interval of parameters), to the Uhlemann et al. PDF in the high density range.

**Key words:** galaxies: clusters: general – dark matter – large-scale structure of Universe.

## 1 INTRODUCTION

The recognition that the Universe contains large-scale structures involved a long process of discovery, from initial speculations to physical theories, guided by increasingly refined data from observational surveys (see a historical account in e.g. Saslaw 2000). These structures, composed of smaller gravitational units (clusters of galaxies, galaxies, etc.), probably formed from small initial fluctuations in the early Universe, and evolved through gravitational instability in an expanding space–time. Observations such as the cosmic microwave background radiation and baryon acoustic oscillation signals (Planck Collaboration XIII 2016), high-redshift Type Ia supernovae (Riess et al. 1998; Perlmutter et al. 1999), etc. converged to a spatially flat $\Lambda$ cold dark matter ($\Lambda$CDM) cosmological model, which however has recently been under significant tension due to revealing inconsistencies (e.g. Garcia-Quintero et al. 2019).

Our current understanding indicates that galaxies (and other observables such as quasars, line intensities maps, diffuse backgrounds, etc.) are only tracers of the large-scale structures, with the mass component being dominated by dark matter (DM), evolving in an accelerated expanding background (due to some form of dark energy). In this $\Lambda$CDM cosmology, large-scale structures are formed hierarchically, in which DM haloes are assembled from the merging of smaller structures. The relation between the distribution of tracers and the underlying distribution of (total) matter, termed 'bias' (Weinberg et al. 2004; Desjacques, Jeong & Schmidt 2018), its nature and how it evolves in time, is of fundamental importance, not only for the understanding and characterization of large-scale structures, but also as tests for cosmological models.

The study of large-scale structures and its non-linear evolution also relies on statistical tools such as the power spectra and correlation functions (Peebles 1980; Bernardeau et al. 2002). Another approach is given by the counts-in-cells (CiC) probability distribution function (PDF; Saslaw 1985, 2000; Efstathiou et al. 1990; Szapudi 1998), in which discrete objects such as galaxies are counted inside cells of fixed size and shape in an ensemble. The CiC can also be expressed as an appropriately smoothed density-in-cells distribution, and it can be derived from fundamental theories of gravitational clustering and/or estimated from survey data (e.g. Uhlemann et al. 2016, 2018; Salvador et al. 2019).

The gravitational quasi-equilibrium distribution (GQED) describes the CiC statistics of $N$-point masses in an expanding universe. The GQED was first derived by Saslaw & Hamilton (1984), leading to several subsequent studies and refinements (summarized in the book by Saslaw 2000, following an earlier foundational exposition in Saslaw 1985). In GQED, the pairwise properties of the gravitational potential are consistently introduced into a gravithermodynamical theory, under certain hypotheses, but it has also been subsequently derived from statistical mechanics principles by Ahmad, Saslaw & Bhat (2002). Another CiC PDF of interest is the negative binomial distribution (NBD), first proposed in a cosmological context by Carruthers & Duong-van (1983), and subsequently explored in Elizalde & Gaztanaga (1992) (see also e.g. Yang & Saslaw 2011; Hurtado-Gil et al. 2017; Wen, Kemball & Saslaw 2020).

The Illustris The Next Generation (TNG) project[1] presents a prominent series of cosmological simulations (Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2018, 2019; Pillepich et al. 2018; Springel et al. 2018), suitable for the study of galaxy formation

★ E-mail: ccordulad@gmail.com

[1] https://www.tng-project.org/

and evolution, following the coupled dynamics of baryons and DM through a state-of-the-art magnetohydrodynamical numerical code (AREPO; Springel 2010). Dark-only counterparts to the full runs are also available, with related studies on the large-scale clustering and bias (e.g. Springel et al. 2018; Martizzi et al. 2019, 2020; Montero-Dorta et al. 2020).

In this work, we investigated the one-point statistics of subhaloes in the publicly available IllustrisTNG simulations (full and dark-only versions of the TNG100-1, TNG100-3, and TNG300-3 runs), in terms of the CiC statistics (the GQED and the NBD), and in terms of the density-in-cells statistics. In the latter case, we derived a density-in-cell PDF for the GQED model, obtained from its CiC counterpart under simplified assumptions (hereafter denoted by 'dGQED PDF' $\equiv \mathcal{P}_{\rm GQED}$, the 'd' prefix referring to density-in-cells). This model was compared to the DM density-in-cells PDF proposed by Uhlemann et al. (hereafter denoted 'UHL PDF' $\equiv \mathcal{P}_{\rm Uhl}$; Uhlemann et al. 2016, 2018). We used the notation $\mu \equiv \log(\rho) \equiv \ln(\rho)$ for the natural logarithm of the density-in-cells. Note that we followed Uhlemann et al. (2018) in the presentation of the corresponding PDFs, in terms of the decimal logarithmic scale, i.e. $\log_{10}(\mathcal{P})$, which are plotted against $\mu$.

This paper is organized as follows. In Section 2, we present our motivation and a few clarifications. This section also presents the selection of the IllustrisTNG runs, a summary of the methods used to extract the one-point statistics from the simulations, the models, and the fitting procedure. In Section 3, we present our results: the CiC distributions and the related GQED and NBD best-fitting parameters, as well as a comparative analysis of the obtained density-in-cells PDFs. In Section 4, we summarize and discuss our results. Appendix A provides a brief study of the residues of the GQED and NBD fits. The cosmology parameters used in this paper are those defined in the IllustrisTNG simulations, namely: $\Omega_{\Lambda,0} = 0.6911$, $\Omega_{\rm m,0} = \Omega_{\rm DM,0} + \Omega_{\rm b,0} = 0.3089$, $\Omega_{\rm b,0} = 0.0486$, $\sigma_8 = 0.8159$, $n_{\rm s} = 0.9667$, and $h = 0.6774$ (Planck Collaboration XIII 2016).

## 2 MOTIVATION AND METHODOLOGY

### 2.1 Motivation

Are the gravithermodynamical assumptions underlying the GQED applicable to the distribution of DM subhaloes in the IllustrisTNG? If so, then, in the case of the full runs, does the complexity of the incorporated physical processes and gravo-magnetohydrodynamics introduce secondary corrections to the GQED? Do other models, not derived (in principle, at least) from gravithermodynamical conditions, provide similar predictions for the one-point statistics, with or without baryonic dissipation? Are additional elements necessary for describing the combined clustering of dark and baryonic matter?

This work attempts to extend the scope of applications of the IllustrisTNG simulations into tests of gravithermodynamical theory, under complex, multicomponent clustering dynamics. Our investigation focuses on the simplest GQED model, i.e. the one introduced by Saslaw & Hamilton (1984) (and subsequently explored in various directions). Before we proceed, we would like to highlight two important points.

(i) *The GQED is not an empirical model*, but based on theoretical considerations. It can be derived consistently either from gravitational thermodynamics or from statistical mechanics (e.g. Ahmad et al. 2002). There is a significant body of literature about the GQED, developed for more than 30 yr, including extensions, as well as numerical and observational results, which were summarized in a book (Saslaw 2000). The original GQED theory assumes that galaxies are $N$-body point masses in an infinite, expanding universe model, but this theoretical idealization has been proved to be adequate to first order in several studies.

(ii) *The simplest form of the GQED, as well as the NBD, has no free parameters*. The GQED modifies the Poisson PDF due to the presence of correlations, and is described by two parameters: the clustering parameter $b$, and the average number of galaxies per cell, $\bar{N}$. The NBD is also similarly described. Clearly, $\bar{N}$ can be obtained directly from the data, but it is also the case for the clustering parameter, which is discussed in more detail in the next section. However, obtaining the GQED from a parametrized fitting is easiest to do, especially because the $b$ parameter would otherwise require the evaluation of equation (2), which is non-trivial in a simulation setting. In any case, it is important to emphasize that *a fitting to the QGED automatically provides an implicit test of the adequacy of the thermodynamical regime pre-supposed in the theory* (cf. Sheth & Saslaw 1996). On the other hand, the NBD model has been argued to be unphysical (Saslaw & Fang 1996), even though it does provide a good fit to the observed CiC PDFs (e.g. Hurtado-Gil et al. 2017; Wen et al. 2020).

Some additional notes that motivate our investigation follows. First, it is important to understand limiting behaviours of the PDF parameters, as they encode information on how the gravitational clustering evolves in terms of spatial regions. For instance, in smaller scales, it is fundamental to search for the effects of mergers on the predictions of the CiC statistics, and on larger scales, indications of non-Poissonian initial conditions. Previous investigations analysed the potential impact of certain extensions on the parameter-free nature of the simplest form of the GQED theory. An important consideration is evolution and scale dependence of the parameters (e.g. Sheth & Saslaw 1996). Therefore, one of the motivations for this work was to analyse the CiC parameters in terms of these dependencies.

Second, the role of the dominant dynamical component (DM) is of great interest, especially whether it is individually associated with galaxies through distinct haloes, or more uniformly distributed throughout space. These properties are important clues for cosmological models. Recently, an analysis of the potential use of the GQED and other models for selecting among different cosmologies has been conducted by Wen et al. (2020). These authors used the 648 Mpc $h^{-1}$ box, Dark Energy Universe Simulations (DEUS; Alimi et al. 2012), to study the CiC PDFs of DM haloes, as a function of scale and redshift, finding nuanced features that nevertheless could be used to observationally distinguish between different cosmologies. In the present case, the IllustrisTNG simulations are all set up with the same cosmological model, therefore our analysis addressed the potential nuances of the one-point statistics of subhaloes under a fixed cosmology, with the emphasis on differences between the presence and absence of baryonic physics.

Finally, the UHL PDF (Uhlemann et al. 2016, 2018) is based on large deviation theory (LDT), which considers the rate at which probabilities of certain events decay, as a characteristic parameter of the problem varies (Bernardeau & Reimberg 2016). Previous key results (e.g. Bernardeau 1992, 1994), concerning the complicated evolution of the large-scale cosmic density field, in which the initial linear fluctuations develop into non-linear modes due to gravitational instability, have been connected to LDT. On the other hand, a possible connection between LDT and gravithermodynamics is a novel problem, hence a comparison of the UHL PDF with the GQED predictions is of great interest.

**Table 1.** Summary of the analysed simulations.

| Simulation (TNG) (1) | Redshift (2) | CiC method (3) | Cell radii (Mpc $h^{-1}$) (4) | $N_{sh}$ (5) | $N_{excl}$ (6) |
|---|---|---|---|---|---|
| 300-3 | 0.00 | Rand (IIS88) | $\mathcal{R}_a = \{10.25, 20.5, 30.75, 41.0\}$, $\mathcal{R}_b = \{6.0, 12.0, 24.0\}$ | 391 144 | 23 535 |
| 300-3 | 0.05 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 396 402 | 23 574 |
| 300-3 | 0.11 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 399 948 | 23 633 |
| 300-3 | 1.00 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 429 194 | 20 056 |
| 300-3 | 2.00 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 384 814 | 11 091 |
| 300-3 | 3.01 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 271 945 | 4425 |
| 300-3 | 4.01 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 156 950 | 1233 |
| 300-3 | 5.00 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 72 591 | 248 |
| 300-3-Dark | 0.00 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 372 177 | 26 627 |
| 300-3-Dark | 0.05 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 373 988 | 26 693 |
| 300-3-Dark | 0.11 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 375 860 | 26 719 |
| 300-3-Dark | 1.00 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 374 473 | 22 318 |
| 300-3-Dark | 2.00 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 307 526 | 12 253 |
| 300-3-Dark | 3.01 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 202 685 | 4879 |
| 300-3-Dark | 4.01 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 108 238 | 1372 |
| 300-3-Dark | 5.00 | Rand | $\mathcal{R}_a, \mathcal{R}_b$ | 47 491 | 293 |
| 300-3-H | 0.00 | Rand (∗) | $\mathcal{R}_b$ | 391 144 | 23 535 |
| 100-3 | 0.00 | Grid (Lei19) | 10.00, 11.25 | 118 820 | 2221 |
| 100-3 | 0.00 | Rand | $\mathcal{R}'_a = \{3.75, 7.5, 11.25, 15.0\}$ | 118 820 | 2221 |
| 100-3 | 5.00 | Rand | $\mathcal{R}'_a$ | 68 031 | 34 |
| 100-3-Dark | 0.00 | Rand | $\mathcal{R}'_a$ | 116 020 | 2333 |
| 100-3-Dark | 5.00 | Rand | $\mathcal{R}'_a$ | 45 700 | 36 |
| 100-1-Mock | 0.00 | Grid | 10.00 | 4371 211 | 4359 784 |
| 100-1-Mock | 0.00 | Rand | $\mathcal{R}'_a$ | 4371 211 | 4359 784 |

*Note.* Columns: (1) simulation label; (2) redshift at which the CiC was applied; (3) applied CiC method; (4) cell radii used for the applied CiC method; (5) total number of candidate subhaloes; and (6) number of excluded subhaloes. The number of cells for the largest selection sphere is 9500 for the Rand method, which was doubled (indicated with an asterisk '∗') in the 300-3-H CiC computation. For the Grid method, a regular grid of 216 000 cells was used.

### 2.2 The IllustrisTNG simulations

The currently publicly available simulations[2] are the TNG300 and TNG100 runs, each with three levels of resolution (with label '3' for the lowest resolution runs). These simulations have sizes of $L_{\rm SIM}(\rm TNG100) = 75$ Mpc $h^{-1}$ and $L_{\rm SIM}(\rm TNG300) = 205$ Mpc $h^{-1}$. The TNG300 runs are more adequate for the statistical analysis of large-scale galaxy clustering. The TNG100 follows the same initial conditions (with updated cosmological parameters) of its predecessor, the Illustris simulation, with volume and resolution lying between the TNG300 and the TNG50 simulations. Hence, the TNG100 is adequate for the study of clustering at a scale of intermediate-mass subhaloes. The TNG300 is more adequate than the TNG100 for the applicability of the GQED, which relies on a gravithermodynamical approach for the description of clustering in the largest structures in the Universe.

We performed a preliminary evaluation to determine the number of CiC computations suitable for our project, given constraints regarding our computational resources and time frame. For the CiC computations, we developed a Python (2019) code to implement two extraction methods, given in Itoh, Inagaki & Saslaw (1988, hereafter IIS88) and Leicht et al. (2019, hereafter Lei19); details of these methods are presented in the next section. We established a grid of runs considering the lowest resolution versions of the both available volumes, namely, the TNG100-3 and TNG300-3 (both full and dark-only runs). In Table 1, we present our grid, resulting in a total of 136 CiC computations. Note that the comoving number of available subhaloes, understood as gravitationally bound structures identified by the SUBFIND algorithm (Nelson et al. 2019), decreases for higher redshifts. Details of Table 1 are explained in the next subsections.

The extraction of the density-in-cells statistics for the ULH PDF is much more computationally demanding than in the case of the CiC statistics. The methodology involved, to be explained in Section 2.4, requires the evaluation of the *DM density field, directly from the simulation snapshot data on particles* (i.e. not from the subhalo catalogue data, as in the case for the CiC statistics). In order to understand this difference, notice that the catalogue file for the TNG100-3-Dark at $z = 0$ is ∼39 MB in size, whereas for the evaluation of the DM density field (as required for the UHL PDF), we need to perform the counting of particles from the simulation snapshot file (∼5.6 GB in size for the same run); hence, the latter procedure results in a significantly higher computational cost. Therefore, for extracting the density-in-cells, we analysed the IllustrisTNG 100-3-Dark run at $z = 0$ only, using 297 non-overlapping spheres of radii 10 Mpc $h^{-1}$, leaving out a margin of 2.5 Mpc $h^{-1}$ from the simulation box limits. The DM density field in each cell was obtained from the DM snapshot particles, whereas the DM density of subhaloes in each cell was computed from the DM subhaloes catalogue. The IllustrisTNG 100-3 full run at $z = 0$ was used to compute the density-in-cells of baryons only, associated with the latter subhaloes.

### 2.3 CiC methods, comoving cell sizes, and proxies for galaxies

The CiC statistics are usually extracted from an ensemble of cells with a certain form and size, in which the number of a given class of objects is counted within each cell. In the present case, the objects are

[2]For complete information, please refer to the IllustrisTNG project site.

the IllustrisTNG subhaloes and the cells are chosen from different sets of comoving spherical cells. The CiC statistics then reflect the underlying PDF, $f_V(N)$, of the number of objects per cell volume $V$, between $N$ and $N + \mathrm{d}N$. The interval or bin $\mathrm{d}N$ should be chosen to produce a smooth enough PDF, and this depends on the data sizes. Here we use the terms CiC PDF and $f_V(N)$ interchangeably.

We considered two methods for obtaining the CiC statistics, given by IIS88 and Lei19, which we label as 'Rand method' and 'Grid method', respectively. The main difference between these methods is the following.

*The Rand method.* This method is explained in IIS88, an early numerical investigation about the adequacy of the GQED for describing of the gravitational $N$-body clustering of 4000 point particles (each representing a galaxy), evolving in an expanding universe (represented by a sphere of comoving unity radius). The gravitational interaction of these particles was followed in comoving coordinates, for different cosmological models (total mass density parameter given by $\Omega_\mathrm{m} = \{0.01, 0.1, 1.0\}$, with cold and warm initial velocity distributions). The CiC method in IIS88 consisted of randomly generating 9500 points within a selection sphere at the centre of the simulation box. This procedure was repeated for each investigated configuration (i.e. for each snapshot at certain scale factors of interest). The radius of this selection sphere depended on the cell radius for the counting procedure. IIS88 used selection spheres of radii $R = \{0.8, 0.7, 0.6, 0.6\}$ for corresponding cell radii given by $r = \{0.1, 0.2, 0.3, 0.4\}$, in which each cell was centred on the randomly generated points. The selection spheres were adopted to avoid boundary effects and to include a sufficient number of particles for the CiC statistics. In other words, a pair $(R_i,\ r_i)$ defines the selection sphere and cell radius for each computation of $f_V(N)$, where $V \equiv V_i$ refers to the cell volume with radius $r_i$.

*The Grid method.* In Lei19, the CiC statistics were obtained for the Illustris TNG100 simulation in the context of 21-cm intensity mapping of neutral hydrogen. The CiC provided the statistics on the mean matter densities of neutral hydrogen, matter, and mass-weighted haloes, in overlapping spheres of comoving radius of $R = 5$ Mpc $h^{-1}$, centred on a regular $128^3$ grid. This produced $\sim$2 million density-in-cells samples, in a redshift range of $z = 1$–5. Their fixed choice of $R$ was determined as a compromise between a sufficiently large cell radius and at the same time a large number of independent cells (hence a small enough cell size) for satisfactory statistics (see section 2 in Lei19 for details).

We performed a preliminary evaluation of the CiC methods using the medium-sized, lowest resolution run, TNG100-3, at $z = 0$. Based on this evaluation, we considered a coarser grid than that in Lei19 due to our computational constraints, i.e. we used a regular $60^3 = 216\,000$ grid, imposing an exclusion margin of 750 kpc to avoid boundary effects. For this preliminary test, we considered two counting cell radii: 10 and 11.25 Mpc $h^{-1}$. For the Rand method, we used the same number of cells (9500) within the selection sphere, applied to the preliminary CiC computations for the TNG100-3 at $z = 0$. Both CiC methods resulted in qualitatively similar outcomes. Computations using both these methods were also performed for the TNG-100-1-Mock. Given the similarity of the results in both CiC methods, we fixed our full, subsequent analysis to the Rand method, using the same CiC parameters as in IIS88.

We summarize two sets of cell radii that we used as follows.

(i) *The $\mathcal{R}_a$-set.* Comoving cell radii as fixed fractions of the simulation size (Rand method; IIS88). We defined a set of cell radii given by $\mathcal{R}_a = \{0.1, 0.2, 0.3, 0.4\} \times R_\mathrm{SIM}$ [Mpc $h^{-1}$], with $R_\mathrm{SIM} = L_\mathrm{SIM}/2$. Each cell centre in set $\mathcal{R}_a$ must be inside a selection radius given, respectively, by $R_\mathrm{sph} = \{0.8, 0.7, 0.6, 0.6\} \times R_\mathrm{SIM}$ [Mpc $h^{-1}$], to avoid boundary effects. This gives different comoving cell sizes for the TNG300-3 and TNG100-3 runs, namely: for TNG300-3, $\mathcal{R}_a = \{10.25, 20.5, 30.75, 41.0\}$ [Mpc $h^{-1}$]; and for TNG100-3, $\mathcal{R}'_a = \{3.75, 7.5, 11.25, 15.0\}$ [Mpc $h^{-1}$].

(ii) *The $\mathcal{R}_b$-set.* Comoving cell radii of $\mathcal{R}_b = \{6.0, 12.0, 24.0\}$ [Mpc $h^{-1}$]. This set was only analysed for the TNG300-3 simulations. This choice of radii is the same as that explored in Yang & Saslaw (2011). The respective selection radii were defined as $R_\mathrm{sph} = \{0.8, 0.7, 0.6\} \times R_\mathrm{SIM}$ [Mpc $h^{-1}$].

We used the SUBFIND object catalogue (cf. Pillepich et al. 2018, and references therein) for the identification of subhaloes the counting procedure. We selected subhaloes (as proxies for galaxies) with attributes restricted to the following criteria.

(i) Subhalo mass range: subhaloes had a minimum mass of $2.5 \times 10^8$ M$_\odot$ $h^{-1}$ (as in Lei19), and a maximum cut-off mass of $\sim 10^{13}$ M$_\odot$ $h^{-1}$.

(ii) Subhalo attribute 'SubhaloFlag' had to be equal to 1 (indicating that the particles composing the subhalo have a cosmic origin).

We do not differentiate between central and satellite galaxies (Pillepich et al. 2018). For the dark-only simulations, we used the same criteria as above. The overall matching between bound structures in the SUBFIND catalogue in both full and dark-only runs is the basis for a direct comparison between their corresponding CiC statistics results. In Table 1, we quote in columns (5) and (6) the total number of initial candidate subhaloes ($N_\mathrm{sh}$) and the number of excluded subhaloes ($N_\mathrm{excl}$) after applying the criteria above.

As mentioned previously, we used different criteria for galaxy proxies for the TNG-100-1 simulation, which was based on the publicly available mock catalogue by Rodriguez-Gomez et al. (2019) (TNG-100-1-Mock). In this case, for selecting admissible subhaloes, we matched the individual subhalo ID's listed in that catalogue to those in the SUBFIND catalogue of the TNG100-1 at $z = 0$. The mock catalogue includes subhaloes with (total) stellar mass greater than $10^{9.5}$ M$_\odot$ and tailored to the Sloan Digital Sky Survey (SDSS; York et al. 2000), as explained in Rodriguez-Gomez et al. (2019). Furthermore, we cut magnitudes in the $r$ band for synthetic objects fainter than 17.77 mag, resulting in 11 427 subhaloes, representing galaxies as would be observed in that band at $z \lesssim 0.05$.

### 2.4 Models and fitting procedure

#### *2.4.1 The GQED (CiC PDF)*

The gravithermodynamical theory leading to the simplest form of the gravitational quasi-equilibrium distribution (GQED) assumes that the galaxy clustering evolves through a series of quasi-equilibrium states due to a cancelling effect of the long range, mean gravitational field, from the expansion of the Universe (Saslaw 2000). The resulting CiC PDF, $f_V(N) \equiv f_\mathrm{GQED}(N)$, representing the probability that a cell of volume $V$ of arbitrary shape contains $N$ galaxies, has a form that modifies the Poisson PDF due to the presence of correlations:

$$f_\mathrm{GQED}(N) = \frac{\bar{N}(1-b)}{N!}[\bar{N}(1-b) + Nb]^{N-1} \times \exp[-\bar{N}(1-b) - Nb], \quad (1)$$

where $\bar{N} = nV$ is the expected average number of galaxies in the given volume $V$, with average density $n$. The aggregation parameter $b$ is related to the degree of clustering of galaxies at a certain state, and represents the average departure from a non-interacting ensemble of cells of a given volume at that state. A more detailed physical interpretation for this parameter, in terms of the average

density and kinetic temperature $T$ of the system, was explored in subsequent developments (e.g. Saslaw & Fang 1996; Ahmad et al. 2002). Numerical studies of the gravitational clustering of galaxies in a range of expanding universe models have qualitatively shown that structures quickly relax to the GQED form of equation (1), and that they subsequently evolve, in general, through a series of quasi-equilibrium states (e.g. IIS88; Itoh, Inagaki & Saslaw 1993; Wen et al. 2020). Each of such states would satisfy equation (1) for a given value of $b$, in other words, the theory admits a time-dependent $b(t)$, which increases slowly from a lower value, as clustering proceeds hierarchically into larger and larger scales with time (Saslaw 1986, 2000).

It is also important to understand how the value of $b$ depends on the counting cell size (e.g. Sheth & Saslaw 1996), and how correlations evolve more rapidly, depending on the spatial scale, as non-linear structures develop first from near-neighbour interactions on smaller scales than in larger ones. This effect is explicitly encoded in the expression for $b$ in terms of the two-point correlation function $\xi$ (e.g. Saslaw & Hamilton 1984; Saslaw 2000):

$$b = -\frac{W}{2K} = \frac{2\pi G m^2 n}{3T}\int_0^\infty \xi(n, T, r) r \, \mathrm{d}r, \tag{2}$$

where $m$ is the average galaxy mass. The virial ratio above actually represents an ensemble average ratio of the gravitational *correlation* energy ($W$) to twice the kinetic energy ($K$) of the *peculiar velocities* of galaxies in an idealized *infinite* universe. Depending on the initial conditions, there could be a characteristic scale above which the correlation function would not contribute to the integral in equation (2), leading to $b \to 0$ (Poisson). For very small scales, with either one or zero galaxies, the distribution function would also tend to a Poissonian one (Saslaw 2000).

### 2.4.2 The NBD (CiC PDF)

The negative binomial distribution (NBD) was initially introduced in cosmology without an underlying physical foundation by Carruthers & Duong-van (1983), as a good approximation to the distribution of Zwicky clusters, and further developed by Elizalde & Gaztanaga (1992). The NBD can be expressed as

$$f_{\mathrm{NBD}}(N) = \frac{\Gamma\left(N + \frac{1}{g}\right)}{\Gamma\left(\frac{1}{g}\right) N!} \frac{\bar{N}^N \left(\frac{1}{g}\right)^{\frac{1}{g}}}{\left(\bar{N} + \frac{1}{g}\right)^{N+\frac{1}{g}}}, \tag{3}$$

where $g$ is the NBD clustering parameter, which is also equivalent to the two-point correlation function, $\xi(V)$ (e.g. Wen et al. 2020).

### 2.4.3 The fitting procedure to the CiC PDFs

We directly fit the GQED and NBD models with the respective $(\bar{N}, b, g)$ as free parameters. The fittings were performed using the CURVE_FIT module in the SciPy library (Jones et al. 2020). We chose the LM optimization method (Levenberg–Marquardt algorithm) in the case of the GQED model fitting, and the TRF (Trust Region Reflective algorithm, with parameter bounds set to 0 and $\infty$) for the case of the NBD model fitting, which has proven necessary for achieving convergence. For handling large values of $N$ for the factorial evaluation, we used the PYTHON module DECIMAL. The fittings were performed on CiC histograms with variable bins widths, d$N$, which mainly depended on the cell size. After a series of tests, the bin widths also had to be adjusted for different redshifts, because of the size and spread of the CiC population as a function of redshift. The bin widths d$N$, as a function of redshift and cell radius, were:

(i) for TNG 100-3 and TNG 100-1-Mock: $\mathrm{d}N_{\mathcal{R}'_a}(z = 0) = \{1, 10, 25, 50\}$ and $\mathrm{d}N_{\mathcal{R}'_a}(z = 5) = \{1, 5, 10, 10\}$;

(ii) for TNG 300-3: $\mathrm{d}N_{\mathcal{R}_a}(z \leq 3.01) = \{1, 10, 30, 50\}$, $\mathrm{d}N_{\mathcal{R}_a}(z > 3.01) = \{1, 1, 1, 2\}$ and $\mathrm{d}N_{\mathcal{R}_b}(z \leq 3.01) = \{1, 3, 10\}$, $\mathrm{d}N_{\mathcal{R}_b}(z > 3.01) = \{1, 1, 1\}$.

In the notation above, bin widths (inside brackets) are listed in the same order as the corresponding sets of cell radii (cf. Section 2.3).

### 2.4.4 The dGQED (density-in-cells PDF)

We derived the matter density-in-cells dGQED PDF, $\mathcal{P}_{\mathrm{GQED}}(\rho)$, corresponding to the GQED, equation (1), under some simplifying assumptions (our ansatz follows from a similar procedure derived for the velocity PDF, $f(v)$, as explained in Saslaw 2000). First, we consider that fluctuations in the counting number $N$ (per fixed cell of volume $V$), among cells in the ensemble, correspond to fluctuations in mass density, so that

$$\rho = \alpha \frac{N\langle m\rangle}{V}, \tag{4}$$

where $\langle m\rangle$ is the average expected value for individual galaxy masses under a uniform Poisson distribution, and $\alpha$ is a factor representing local departures from a uniform distribution. We make the simplifying assumption that $\alpha$ is a constant, given by its average value over the entire ensemble. We then rescale equation (1) from number fluctuations to matter density fluctuations by replacing $N$ with $N\langle m\rangle = \rho V/\alpha$, and the average counting number in the ensemble, $\bar{N}$, with $\bar{N}\langle m\rangle = \bar{\rho} V/\alpha$, where $\bar{\rho}$ is the average matter density-in-cells in the ensemble. Next, we use the identity $N! = \Gamma(N + 1)$, making the continuous replacement for $N$. Finally, we convert $f_{\mathrm{GQED}}(N)\Delta N \mapsto \mathcal{P}_{\mathrm{GQED}}(\rho)\,\mathrm{d}\rho$ using $\Delta N \mapsto \beta\,\mathrm{d}\rho$, with $\beta \equiv V/\alpha$. The resulting PDF is given by

$$\mathcal{P}_{\mathrm{GQED}}(\rho) = \frac{\beta^2\bar{\rho}(1-b)}{\Gamma(\beta\rho+1)} \{\beta\,[\bar{\rho}(1-b) + \rho b]\}^{\beta\rho-1} \times \exp\{-\beta\,[\bar{\rho}(1-b) + \rho b]\} \qquad [\mathrm{dGQED\ PDF}]. \tag{5}$$

The dGQED PDF is presented in Fig. 1, in which we illustrate its behaviour under variations in $\beta$, $\bar{\rho}$, and $b$. In particular, the behaviour of the dGQED in the latter panel (variations in $b$) is qualitatively similar to the GQED CiC PDF, as can be seen in the panels shown in fig. 28.1 of Saslaw (2000).

We also derive an approximate relation between the density-in-cells variance and the $b$ parameter for the dGQED. From equation (26.34) and subsequent discussions in the chapter 28 of Saslaw (2000), the variance of fluctuations for the number of objects in a cell, in the case of the GQED, is given by $\sigma^2 \approx \bar{N}(1-b)^{-2}$ (to order $\bar{N}^{-1/2}$). In our ansatz, we propose $\sigma^2_\rho \approx \bar{\rho}(1-b)^{-2}$. Then, using the approximate correspondence of variances in log-densities and densities, $\mathrm{Var}[\log\rho] \approx \frac{1}{\bar{\rho}^2}\mathrm{Var}[\rho]$ (where the expectation value is $E[\rho] = \bar{\rho}$), we find

$$\sigma^2_{\log\rho} \equiv \sigma^2_{\mu;\mathrm{dGQED}} \approx \frac{1}{\bar{\rho}}\frac{1}{(1-b)^2}. \tag{6}$$

### 2.4.5 The UHL (density-in-cells PDF)

The UHL PDF is based on a bias model, relating the density-in-cells statistics of the DM field, $\rho_{\mathrm{m}}$, and those of the DM subhaloes, $\rho_{\mathrm{sh}}$ (Uhlemann et al. 2016, 2018). Here we use the saddle-point approximation for the integral form of the PDF, as explained in Uhlemann et al. (2016). Up to the mildly non-linear clustering regime

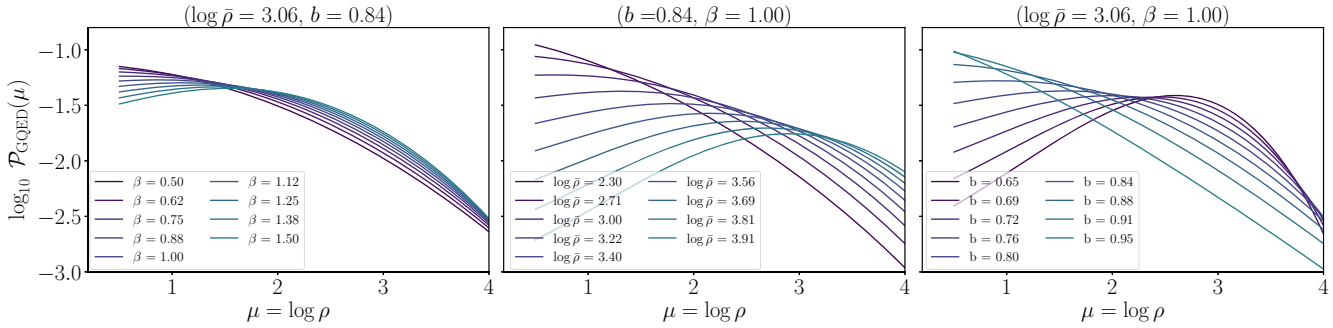


**Figure 1.** Behaviour of equation (5) under variations in $\beta$, $\bar{\rho}$, and $b$ (from left- to right-hand panels, respectively). The titles over each panel indicate the values of the quantities that were held fixed.

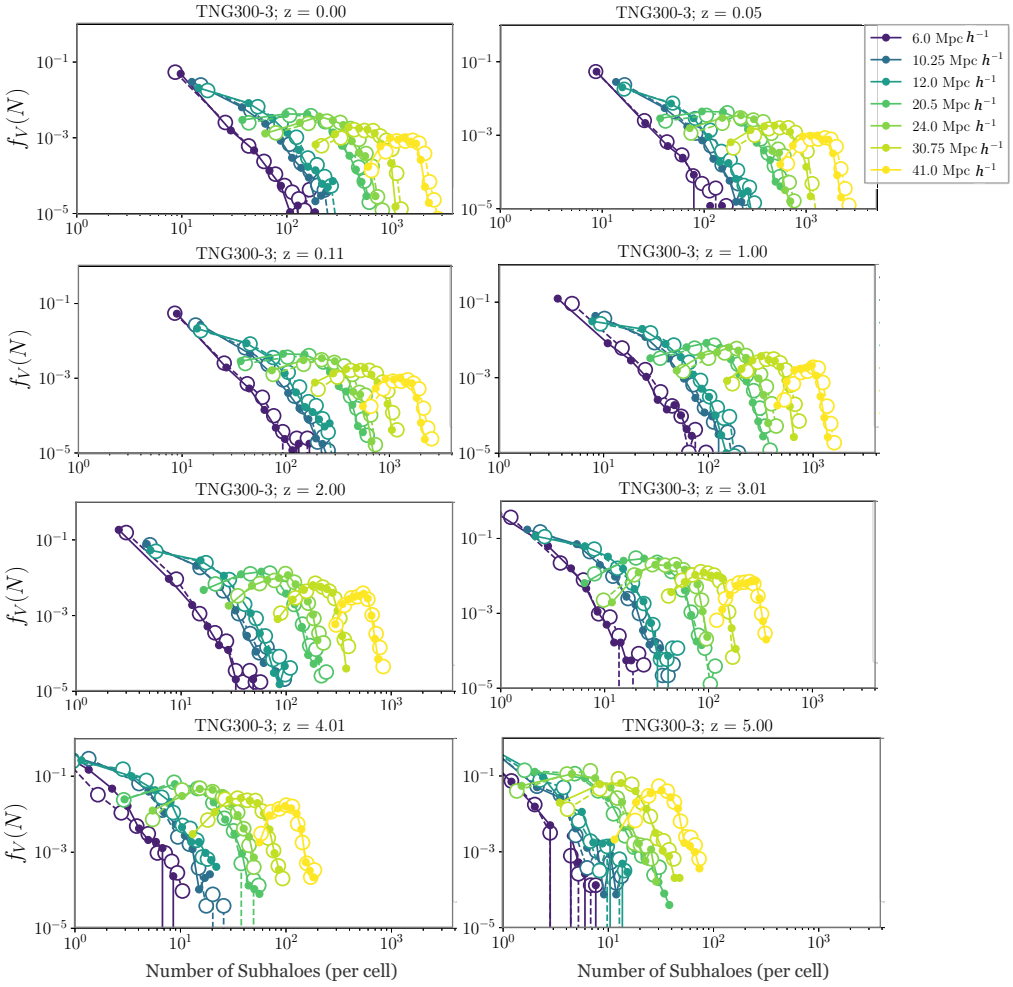


**Figure 2.** Normalized CiC histograms (in log–log scale) for the TNG-300-3 simulations, starting from redshift $z = 0.0$ (top, left) to $z = 5.00$ (bottom, right), for comoving cell sizes indicated in the legend (covering both $\mathcal{R}_a$ and $\mathcal{R}_b$ sets, cf. Table 1). Darker colours are used for smaller radii. The number of bins in this representation is fixed to $N_{\rm bins} = 10$. Symbols are used instead of histogram bars: smaller (filled) circles connected with continuous lines for the full runs; larger (empty) circles connected with dashed lines for the corresponding dark-only runs.

($\sigma_\mu^2 \lesssim 1$), the UHL PDF for $\rho_{\rm m}$ within a sphere of radius $R$ and redshift $z$ is given by

$$\mathcal{P}_R(\rho_{\rm m}) = \sqrt{\frac{\Psi_R''(\rho_{\rm m}) + \Psi_R'(\rho_{\rm m})/\rho_{\rm m}}{2\pi\sigma_\mu^2}}\,\exp\left(-\frac{\Psi_R(\rho_{\rm m})}{\sigma_\mu^2}\right), \tag{7}$$

with

$$\Psi_R(\rho_{\rm m}) = \frac{\tau_{\rm SC}^2(\rho_{\rm m})\sigma_{\rm L}^2(R)}{2\sigma_{\rm L}^2(R\rho_{\rm m}^{1/3})}, \tag{8}$$

$$\tau_{\rm SC}(\rho_{\rm m}) = \nu\left(1 - \rho_{\rm m}^{-1/\nu}\right), \tag{9}$$

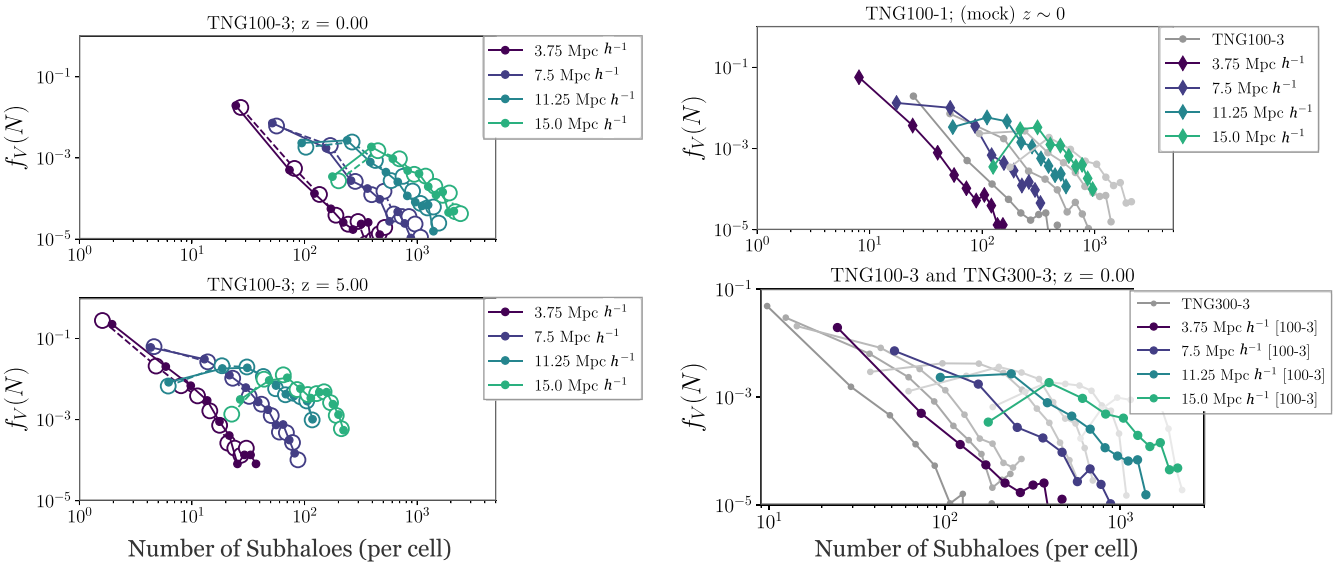


**Figure 3.** Normalized CiC histograms (in log–log scale) for the TNG-100-3 simulations. Left-hand panels: redshift $z = 0.0$ (top) and $z = 5.00$ (bottom). Right-hand panels: in the top panel, the TNG-100-1-Mock run, with diamond symbols, whereas filled circles in grey-scale represent the results for TNG-100-3 full at $z = 0.0$; in the bottom panel, the TNG100-3 full distributions are compared with those of the TNG300-3 full runs (both at $z = 0.0$), showing the effects of volume size.

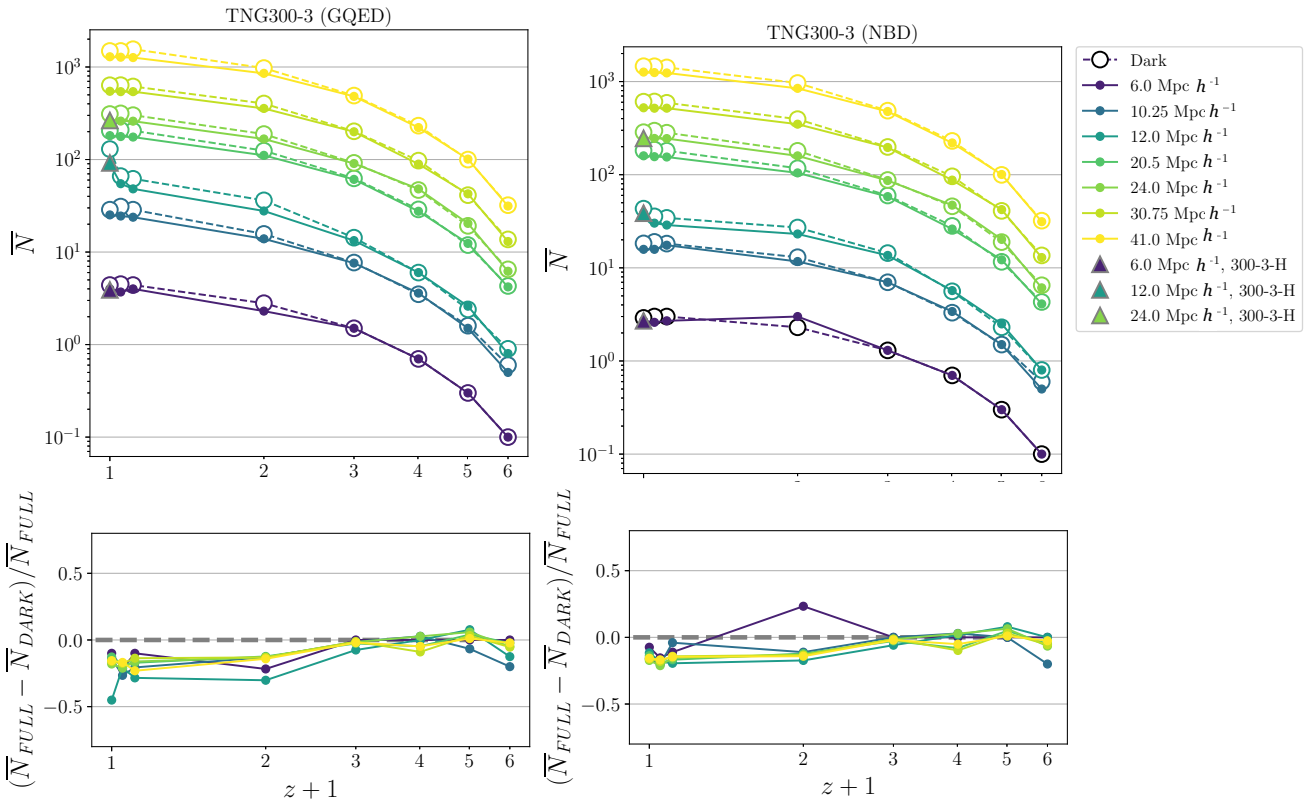


**Figure 4.** Results for the TNG300-3 simulations, redshifts $z = \{0.00, 0.05, 0.11, 1.00, 2.00, 3.01, 4.01, 5.00\}$, for various comoving cell radius (darker colours for smaller radii, see legend). Filled circles connected with continuous lines refer to the full simulation runs; larger, empty circles connected with dashed lines refer to the respective dark-only simulation runs. The TNG300-3-H (full) runs are represented by triangles at $z = 0$. Top panels: behaviour of the average number of subhaloes per cell, $\bar{N}$, as a function of $z$, obtained from the fit to the GQED (left-hand panel) and the NBD (right-hand panel) models. Bottom panels: fractional difference between full and dark-only results, relatively to the full runs (GQED: left-hand panel; NBD: right-hand panel).

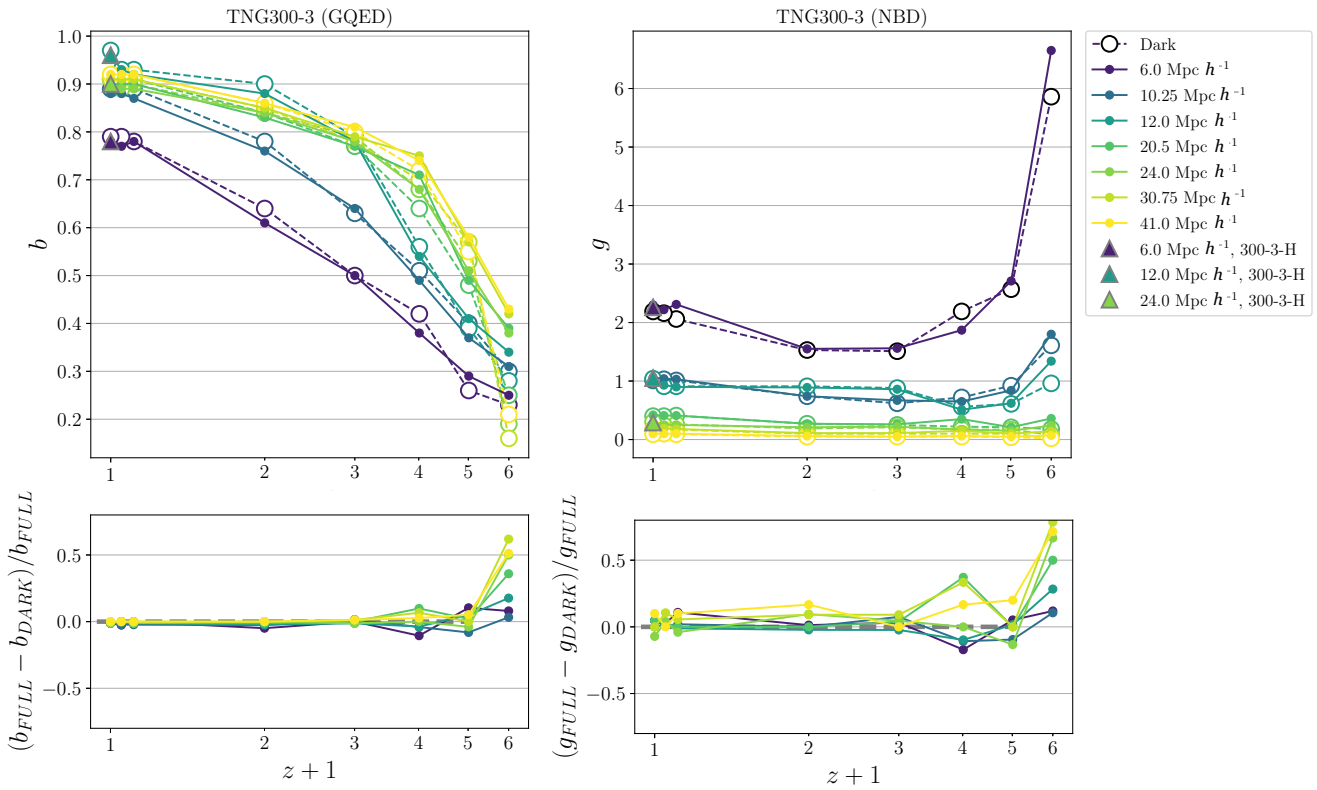


**Figure 5.** Results for the TNG300-3 simulations, redshifts $z = \{0.00, 0.05, 0.11, 1.00, 2.00, 3.01, 4.01, 5.00\}$. Top panels: behaviour of the clustering parameter as a function of $z$, obtained from the CiC PDF fit to the GQED ($b$ parameter, left-hand panel) and NBD ($g$ parameter, right-hand panel) models, for various comoving cell radius, as indicated in the legend. Bottom panels: fractional difference between full and dark-only results, relatively to the full runs (GQED: left-hand panel; NBD: right-hand panel). Symbols are the same as in the previous figure.

where we used $\nu = 21/13$; the prime in equation (7) denotes a derivative with respect to $\rho_\mathrm{m}$, $\sigma_\mu$ is the non-linear variance of the corresponding log-density ($\mu = \log\rho$), $\tau_\mathrm{SC}$ is the linear density contrast averaged within the Lagrangian radius (see details in Uhlemann et al. 2018), and $\sigma_\mathrm{L}$ is the linear variance of the density field on a scale $R$. For the latter, we used the SIGMA module of the COLOSSUS PYTHON toolkit for cosmological calculations (Diemer 2018), giving $\sigma_\mathrm{L}^2 = 0.486$. We fit the data to the quadratic bias model in the log-densities, given by

$$\mu_\mathrm{m} = b_0 + b_1\mu_\mathrm{sh} + b_2\mu_\mathrm{sh}^2. \tag{10}$$

The best-fitting parameters provided the mean relation $\mu_\mathrm{m}(\mu_\mathrm{sh})$ of the bias model, and the subhalo PDF $\mathcal{P}_\mathrm{sh}$ was obtained from the DM PDF $\mathcal{P}_\mathrm{m}$ (equation 7) by conservation of probability:

$$\mathcal{P}_\mathrm{Uhl} \equiv \mathcal{P}_\mathrm{sh}(\rho_\mathrm{sh}) = \mathcal{P}_\mathrm{m}[\rho_\mathrm{m}(\rho_\mathrm{sh})] \left|\frac{\mathrm{d}\rho_\mathrm{m}}{\mathrm{d}\rho_\mathrm{sh}}\right| \qquad \text{[UHL PDF]}. \tag{11}$$

We also used the large density tail ($\rho \gg 1$) of the UHL PDF, as given in Uhlemann et al. (2016) (their equation 27):

$$\mathcal{P}_\mathrm{tail}(\rho_\mathrm{m}) \to \frac{(n_\sigma+3)\nu}{6\sqrt{\pi\sigma_\mu^2}} \exp\left[-\frac{\nu^2(\rho_\mathrm{m}^{\frac{1}{\nu}}-1)^2\rho_\mathrm{m}^{\frac{n_\sigma+3}{3}-\frac{2}{\nu}}}{2\sigma_\mu^2}\right] \rho_\mathrm{m}^{\frac{n_\sigma-3}{6}}, \tag{12}$$

where we used the notation $n_\sigma$ for the index of a power-law initial power spectrum, occurring in their equation (13); namely, the variance of the field fluctuation within a sphere of radius $R$ follows the relation

$$\sigma^2(R) = \sigma^2(R_\mathrm{p})\left(R/R_\mathrm{p}\right)^{-(n_\sigma+3)}, \tag{13}$$

where $R_\mathrm{p}$ is a pivot scale (see their paper for details).

We developed a PYTHON code to implement these calculations. The positivity condition involving the derivatives of the $\Psi$ function in equation (7) was met in the ranges analysed; hence the saddle-point approximation was adequate. We checked the validity of our code by testing against data ranges and parameters found in Uhlemann et al. (2018). For a set of UHL PDFs computed with a different code (LSSFAST), for different variances and radii, see Codis et al. (2016).

## 3 RESULTS

### 3.1 Qualitative presentation of the CiC distributions

Following a similar display as in Wen et al. (2020), we present all our normalized CiC PDFs (in log–log scale) in Figs 2 and 3 for the TNG300-3 and TNG100-3 runs, respectively. We also present the TNG-100-1-Mock results (Fig. 3, top right-hand panel) with TNG100-3 full results together in the same panel for direct comparison. Finally, we present in Fig. 3, bottom right-hand panel, a comparison between TNG300-3 and TNG100-3, for the full runs at $z = 0$. The resulting CiC distributions agree with previous numerical results in the literature, even though being purely gravitational ones (e.g. IIS88; Itoh et al. 1993; Yang & Saslaw 2011; Hurtado-Gil et al. 2017; Wen et al. 2020).

Full and dark-only simulations follow remarkably similar CiC distributions, in both the TNG300-3 and TNG100-3 runs. There are, however, differences that can be finely traced by a quantitative analysis of the best-fitting parameters, to be discussed in Section 3.2.

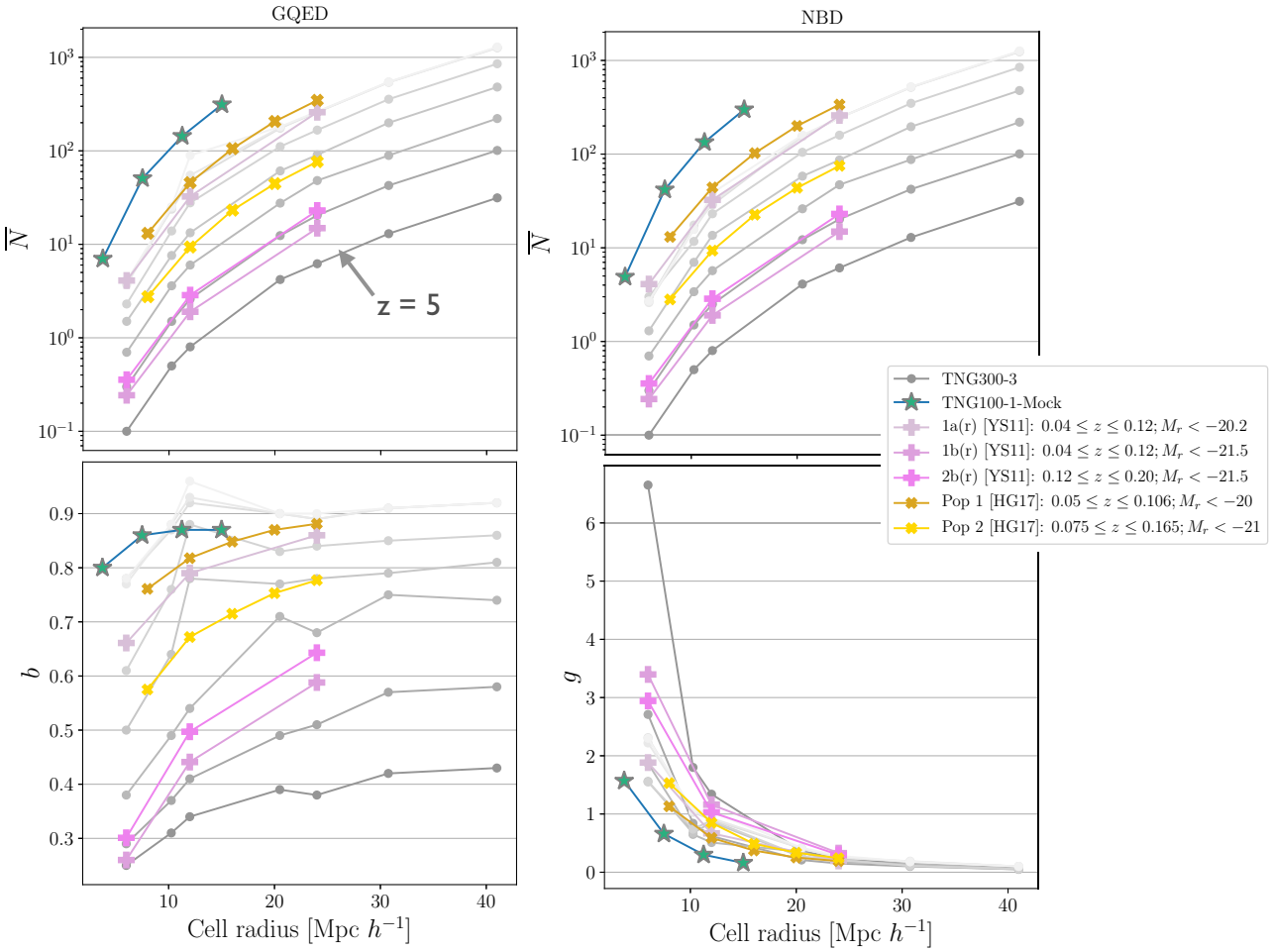


**Figure 6.** Behaviour of the best-fitting parameters for the TNG300-3 runs as a function of comoving cell size; full circles connected with thin lines are shown with grey-scale tones such that the darker tones represent greater redshifts ($z = 5$ results are indicated for reference). Also shown are the TNG100-1-Mock results, as well as those for the corresponding best-fitting values for the data in YS11, in terms of the 1a(r), 1b(r), and 2b(r) samples, are indicated. Also shown are results for the data in HG17, for the Population 1 (Pop 1) and Population 2 (Pop 2) samples. Redshift and $M_r$ magnitudes ranges for those data are indicated in the legend. Top panels: behaviour of the average number of subhaloes per cell, $\bar{N}$, as a function of comoving cell size, obtained from the CiC PDF fit to the GQED (left) and NBD (right) models. Bottom panels: behaviour of the best-fitting clustering parameters, as a function of comoving cell size: $b$ parameter (GQED, left) and $g$ parameter (NBD, right).

Both simulations volumes cover (approximately) similar $N$ ranges in their CiC PDFs, even though the comoving cell sizes are different. All CiC distributions tend to gradually spread towards $z = 0$ as more subhaloes are formed and the gravitational clustering evolves. The height of the CiC PDF peak decreases for larger cell sizes due to spreading and normalization, even considering a larger number of samplings. The height of the CiC PDF peak decreases for lower redshifts, as the distribution spreads into a wider range of $N$. Note, however, that in Wen et al. (2020) (cf. their fig. 2), the trend goes in opposite direction: the peak height decreases at higher redshifts. as the distribution spreads into a wider range of $N$ for *physical* cell sizes, as explained in their paper. The reason for this difference appears to be due to a different measurement criterion for the counting cells. In Wen et al. (2020), the authors use physical cell sizes, whereas we use comoving cell sizes. In the former case, the cell encloses a larger volume at higher redshifts, whereas we follow the same comoving volume.

Our statistics are necessarily poorer for higher redshifts and smaller cell sizes, as the total number of subhaloes available for the comoving counting (gravitationally bound structures) is less than that at $z = 0$ (cf. Table 1, column $N_{\rm sh}$). An analysis of the cell sampling showed that, considering the TNG300-3-H case, doubling of the number of counting cells leads only to a linear increase on the resulting $N_{\rm samples}$ (at least for $z = 0$). We conclude that the CiC statistics for the smallest cell size would require at least an order of magnitude increase in the number of counting cells, so our results for the smallest cell size in both TNG100-3 and TNG300-3 runs should be considered less certain. In Wen et al. (2020), the CiC distribution was found to be less smooth in larger cells at higher redshifts; we attribute this opposite effect given the physical (theirs) versus comoving (ours) counting cell method.

### 3.2 Quantitative analysis of the best-fitting parameters

#### *3.2.1 CiC best-fitting parameters*

In Fig. 4, we show the best-fitting average number of subhaloes per cell as a function of redshift, for the TNG300-3 runs (GQED: left-hand panel; NBD: right-hand panel). The full and dark-only runs follow closely the same overall behaviour of this parameter. In order to see more clearly distinctions between the full and dark-only results,

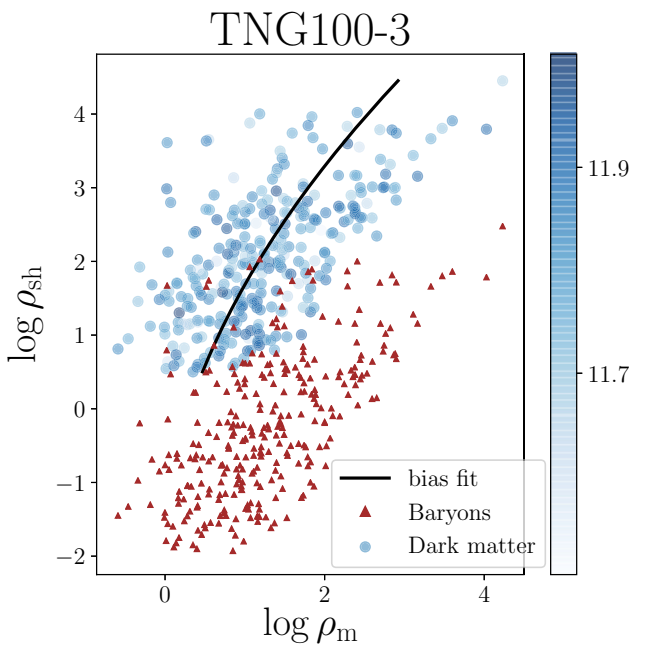


**Figure 7.** Scatter plot of density-in-cells for the dark matter (DM) field ($\rho_m$) versus density-in-cells for the subhaloes ($\rho_{sh}$), extracted from the IllustrisTNG 100-3-Dark run at $z = 0$, with a fixed cell radius of $R = 10$ Mpc $h^{-1}$. The fitting of the bias model, equation (10), is shown. For comparison, the corresponding scatter plot for the density-in-cells of baryons only, within the same selected haloes, is shown (extracted from the corresponding IllustrisTNG 100-3 full run). The vertical colour bar indicates the values of $\log_{10}\langle M_s\rangle$ (the average mass in the respective cells).

we show in the bottom panels of Fig. 4 the fractional difference between these runs, relatively to the full runs, as a function of redshift, for the given cell sizes. At lower redshifts, the $\bar{N}$ parameter tends to be systematically higher for the dark-only runs. This effect is small ($\lesssim$5 per cent) but clearer in the GQED case.

We present in Fig. 5 the results for the TNG300-3 runs in terms of the clustering parameters $b$ (GQED; left-hand panel) and $g$ (NBD; right-hand panel), as functions of redshift and cell size, in the same format as in Fig. 4. For the GQED, we see a clear trend, for both full and dark-only runs and for all cell sizes, showing smaller values of $b$ at higher redshifts to larger values of $b$ at lower redshifts. Smaller cells (at fixed redshift) tend to show smaller values of $b$. For the NBD, $g$ values are generally larger for smaller cell sizes. The bottom panels of Fig. 5 show the fractional difference between full and dark-only $b$ (GQED) and $g$ (NBD) results, relatively to the full runs, as a function of redshift, for the given cell sizes. Overall, both full and dark-only runs present a remarkably similar evolution of their respective clustering parameters at redshifts $z \lesssim 2$, with greater values of $b$ and $g$, for the full runs and for larger cell sizes at $z = 5$.

For the TNG100-3, we have only analysed the $z = \{0.00, 5.00\}$ runs, and for brevity we omit related figures. The best-fitting parameters follow in this case the same overall trends in terms of comoving cell size as in the TNG300-3 runs, at the $z = 0$ and $z = 5$ values. The main differences between these simulation volumes are the following. For the $\bar{N}$ parameter, in both GQED and NBD, the TNG100-3 values at $z = 5$ are higher than those of the TNG300-3 runs, but the increase of this parameter towards the values at $z = 0$ is less than one order of magnitude, i.e. a lower relative increase than in the TNG300-3 runs. The $b$ parameter shows higher values at $z = 5$ as compared to the TNG300-3 runs, but converges to those of the TNG300-3 runs at $z = 0$.

### 3.2.2 CiC: IllustrisTNG and observations

We compare our results with those obtained by Yang & Saslaw (2011, hereafter YS11) and Hurtado-Gil et al. (2017, hereafter HG17). In YS11, the galaxy catalogue is a flux-limited ($r < 17.6$) subsample taken from SDSS Data Release 7 (DR7; Abazajian et al. 2009), with additional absolute magnitude cuts, resulting in three subsamples within two redshift ranges, namely: 1a(r): $0.04 \leq z \leq 0.12$, $M_r < -20.2$; 1b(r): $0.04 \leq z \leq 0.12$, $M_r < -21.5$; 2b(r): $0.12 \leq z \leq 0.20$, $M_r < -21.5$. The lower cut at $z \leq 0.04$ excludes the Coma and Virgo clusters; the subsample follows the Hubble flow. On the other hand, the higher redshift range includes the SDSS great wall, allowing a comparison of potential differences between both ranges. In HG17, the data were also based on the DR7, with galaxy catalogue provided by the New York University Value-Added Galaxy Catalog (NYU-VAGC; Blanton et al. 2005). They also used the LasDamas simulation catalogue (McBride, Berlind & Scoccimarro 2011) to estimate the uncertainties in the resulting CiC distribution. Their selected samples are given by two populations: Population 1 (Pop 1): $0.050 \leq z \leq 0.106$, $M_r < -20.0$; Population 2 (Pop 2): $0.075 \leq z \leq 0.165$, $M_r < -21.0$; which are placed roughly within samples 1a(r) and 2b(r) from YS11, respectively.

In Fig. 6, we present the behaviour of the best-fitting parameters $\bar{N}$, $b$, and $g$ for the TNG300-3 runs, as a function of comoving cell size, for all redshifts analysed (darker grey-scale tones indicate greater redshifts, with the $z = 5$ results indicated for reference). Also shown are the TNG100-1-Mock results, and the results from YS11 and HG17. All samples (observational and simulated) follow similar behaviours in terms of cell size, for both GQED and NBD models, only differing in terms of relative amplitudes in their best-fitting parameters. Trends in $b$ and $g$ are more heterogeneous across the samples and simulation runs than those found in $\bar{N}$.

### 3.2.3 Density-in-cells

Here we report the results of the density-in-cells statistics for the IllustrisTNG 100-3-Dark run at $z = 0$. In Fig. 7, we present the fitting of the bias model, equation (10), over the scatter plot of density-in-cells for the DM field (obtained from the particle snapshot data) versus density-in-cells for subhaloes (obtained from the subhaloes catalogue). We also show, for comparison, the corresponding scatter plot for the density-in-cells of baryons only, within the same selected haloes (obtained from the IllustrisTNG 100-3 full run at $z = 0$). The bias best-fitting parameters were $b_0 = 0.297$, $b_1 = 0.312$, and $b_2 = 0.062$. The measured non-linear variances in the log-densities were $\sigma^2_{\mu,m} = 0.669$ and $\sigma^2_{\mu,sh} = 0.755$.

In Fig. 8 (left-hand panel), we present the normalized histogram of the density-in-cells of DM subhaloes, representing the probability density function at each bin (the counting was normalized in order to give a unit integral over the range). Based on the bias best-fitting parameters, the predicted PDF for the log-densities of subhaloes, $\mathcal{P}_{Uhl}$ (equation 11; see also equation 7), was obtained. We also plot the corresponding high-density tail of the PDF (equation 12); in this case, a normalization of the curve was applied to achieve a good fit for the histogram at $\mu_{sh} > 2.0$, with $n_\sigma = -2.4$. The dGQED PDF, $\mathcal{P}_{GQED}$ (equation 5), was also plotted in Fig. 8 for comparison. In this case, we inverted equation (6) for the expected $b$ parameter, using

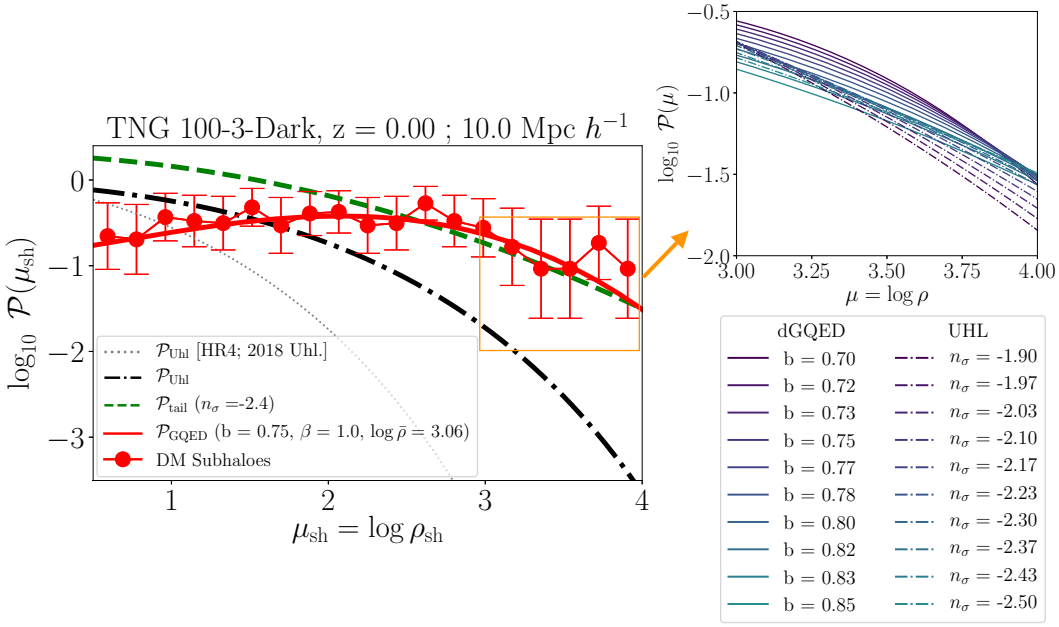


**Figure 8.** Left-hand panel: normalized histogram of the density-in-cells of dark matter (DM) subhaloes in the IllustrisTNG 100-3-Dark run at $z = 0$ (error bars are given by the normalized bin counts, weighted by the non-linear rms variance of the log-density of subhaloes). The PDF curves for the log-densities of subhaloes (in decimal logarithmic scale) are predictions from: (i) the UHL PDF (equation 11; dot–dashed line), based on the measured DM density field variance and best-fitting bias parameters (see Fig. 7); (ii) the high-density tail of the UHL PDF (equation 12, normalized; dashed line); and (iii) the dGQED PDF (equation 5, normalized; continuous line). Also shown is the $\mathcal{P}_{\rm Uhl}$ curve for the Horizon Run 4 (thin dotted line) at the same $z$ and $R$ values (Uhlemann et al. 2018). Right-hand panel: detail of the predictions of the PDF models in the high-density region, as indicated by the box and arrow in the left-hand panel. We show a comparison between the high-density tail of the UHL PDF (dot–dashed lines), for a range of $n_\sigma$, and the dGQED PDF (continuous lines), for a range of $b$ (with $\log\bar{\rho} = 3.06$, $\beta = 1.0$ fixed).

$\sigma^2_{\mu,\rm sh} = 0.755$, giving a good fit for $\log\bar{\rho} = 3.06$; a normalization of the curve was also applied.

As a reference, we also show in Fig. 8 (left-hand panel) the approximate $\mathcal{P}_{\rm Uhl}$ curve for the Horizon Run 4 (HR4) simulation (Kim et al. 2015), obtained in Uhlemann et al. (2018), also for $z = 0$ and $R = 10$ Mpc $h^{-1}$. The computed $\mathcal{P}_{\rm Uhl}$ curves for the IllustrisTNG and HR4 evidently differ, given the differences in those simulations, such as cosmological model, box size, range of resolved halo masses, etc. The HR4 is a ΛCDM cosmological *N*-body simulation, set with the 5-year *Wilkinson Microwave Anisotropy Probe* (*WMAP*5) cosmology parameters, differing from the *Planck*-16 values used in the IllustrisTNG. Also, the HR4 is 3.15 Gpc $h^{-1}$ box, much larger than the IllustrisTNG 100. The selected haloes in the HR4, used to validate the ULH PDF, have masses ranging from $2.7 \times 10^{11}$ to $4.2 \times 10^{15}$ $M_\odot$ $h^{-1}$, to be contrasted with our selection of subhalo masses, from $2.5 \times 10^8$ to $\sim 10^{13}$ $M_\odot$ $h^{-1}$. Also as a reference, for a fitting to the UHL PDF in the IllustrisTNG 100 at $R = 5$ Mpc $h^{-1}$ and $z = \{1, 2, 3, 4, 5\}$; see Lei19.

Our results indicate that the (appropriately normalized) $\mathcal{P}_{\rm tail}$ (equation 12) of the UHL PDF fits well the high density range of the extracted density-in-cells statistic, but not the saddle-point approximation $\mathcal{P}_{\rm Uhl}$ (equation 11), which has proven valid up to variances of the DM log-density of $\sigma^2_\mu \sim 0.5$ (Uhlemann et al. 2016). The variances in the log-densities obtained in our sampling ($\sigma^2_{\mu,\rm m} = 0.669$, $\sigma^2_{\mu,\rm sh} = 0.755$) are somewhat above that limit, which might explain the mismatch. Nevertheless, the UHL PDF is compatible with the data for the limited intermediate range of $1 \lesssim \log\rho_{\rm sh} \lesssim 2$. The very low density range of the UHL PDF predicts a slight excess over the extracted data. Note that the high-density tail required a value of $n_\sigma = -2.4$, which is at the limit to avoid the criticality of the decay-rate function (see fig. 1 in Uhlemann et al. 2016); nevertheless the fitting was adequate in that range. On the other hand, the (appropriately normalized) dGQED PDF fitted very well the entire range of extracted density-in-cells.

We point out that, as the clustering parameter $b$ increases (for a fixed $\bar{\rho}$; cf. the right-hand panel in Fig. 1), the dGQED PDF becomes flatter and skews in comparison to a Poissonian PDF. A qualitatively similar behaviour is seen in the UHL PDF, in terms of the variances of the density-in-cells statistics, as can be seen in fig. A1 of Codis et al. (2016), in which the PDF deviates from a Gaussian (at low variances) to a much skewed distribution towards the high density range (at larger variances). This effect is also a function of time, as the initial PDF becomes gradually more skewed at lower redshifts, as voids increase in extent and density peaks increase in amplitude (clustering increases) via accretion of matter.

Interestingly, the high-density tail of the UHL PDF approaches well the dGQED curve in that range. We investigated this proximity by plotting together both (normalized) PDFs in Fig. 8 (right-hand panel), in the high-density region, for a range of $n_\sigma$ in the case of the UHL PDF (equation 12) and for a range of $b$ in the case of the dGQED PDF, fixing $\log\bar{\rho} = 3.06$, $\beta = 1.0$ (equation 5). Clearly, the differences in the predictions can reach a low percentage or even subpercentage levels in that density range.

## 4 SUMMARY AND CONCLUSIONS

We analysed the compatibility of the one-point statistics of subhaloes in the IllustrisTNG simulations with predictions of four models: the GQED, the NBD, the dGQED, and the UHL PDFs. We extracted

136 CiC samples from the IllustrisTNG 300-3, 100-3, and 100-1 (full and dark-only) runs, in a range of cell sizes and redshifts. For the density-in-cells extraction, we used the IllustrisTNG 100-3-Dark at $z = 0$, at a fixed cell radius of $R = 10$ Mpc $h^{-1}$.

We found that both the full and dark-only runs follow similar GQED and NBD CiC PDF forms. The two simulation boxes cover similar ranges in CiC number counts. Comparing our results in the literature, we found similar scaling and evolutionary trends in all samples, up to factors in the amplitude of parameter values in the case of observational data, possibly regulated by different magnitude cut-offs. Despite the similarity of the CiC PDFs, we found measurable differences between full and dark-only runs, leading to trends in the fitting parameters, which might be relevant for the understanding of bias in terms of gravithermodynamical predictions. For example, the clustering parameter $b$ in the full runs converged approximately to the dark-only runs at lower redshifts, but then $\bar{N}$ tended to become smaller relatively to the dark-only runs. This could be an indication that subhaloes in the full runs were merging inside common DM haloes more efficiently than in the dark-only runs.

The UHL PDF in the saddle-point approximation was compatible with an intermediate range of densities, with the (normalized) high-density tail separately giving a good fit. The (normalized) dGQED PDF fitted very well the entire range of extracted data. Interestingly, we found that, after normalization, dGQED and UHL PDFs in the high density range approximated each other to sub-percentage levels for different parameters. This sector in the PDFs corresponding to rare events, namely, large density fluctuations, is important to constraint the dynamics and cosmology from the initial conditions in the density field to the final distribution (e.g. Codis et al. 2016; Uhlemann et al. 2018; Lei19; Wen et al. 2020).

Our work attempted to extend the scope of applications of the IllustrisTNG simulations into tests of gravithermodynamical theory under complex, multicomponent physics, and to compare its performance against other predictions. For the most part, we found that the gravitational quasi-equilibrium thermodynamical assumptions still hold in the presence of baryonic physics, with residues increasing for the smallest cell size. Given the open problem concerning the physical basis of the NBD (Saslaw & Fang 1996), the meaning of adequate fittings to this PDF is unclear as to the level of its flexibility against variability of the data. The qualitative differences and similarities encoded in the analysed one-point PDFs could enable an increased understanding of their common elements of validity, or assumptions to be discarded or modified. For instance, our results suggest that a connection between LDT and gravithermodynamics, in the case of high-density events, could lead to novel insights. Such a development may be relevant for providing specific predictions for future large galaxy surveys.

## ACKNOWLEDGEMENTS

CCD thanks Dr Hugo V. Capelato for his encouragement during the development of this project. CCD thanks Dr Cora Uhlemann for helpful clarifications. CCD also thanks the referee for corrections and feedback, which greatly improved this work.

## DATA AVAILABILITY

The data underlying this paper will be shared on reasonable request to the corresponding author.

## REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Ahmad F., Saslaw W. C., Bhat N. I., 2002, ApJ, 571, 576
Alimi J.-M. et al., 2012, preprint (arXiv:1206.2838)
Bernardeau F., 1992, ApJ, 392, 1
Bernardeau F., 1994, A&A, 291, 697
Bernardeau F., Reimberg P., 2016, Phys. Rev. D, 94, 063520
Bernardeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, Phys. Rep., 367, 1
Blanton M. R. et al., 2005, AJ, 129, 2562
Carruthers P., Duong-van M., 1983, Phys. Lett. B, 131, 116
Codis S., Pichon C., Bernardeau F., Uhlemann C., Prunet S., 2016, MNRAS, 460, 1549
Desjacques V., Jeong D., Schmidt F., 2018, Phys. Rep., 733, 1
Diemer B., 2018, ApJS, 239, 35
Efstathiou G., Kaiser N., Saunders W., Lawrence A., Rowan-Robinson M., Ellis R. S., Frenk C. S., 1990, MNRAS, 247, 10p
Elizalde E., Gaztanaga E., 1992, MNRAS, 254, 247
Garcia-Quintero C., Ishak M., Fox L., Lin W., 2019, Phys. Rev. D, 100, 123538
Hurtado-Gil L., Martínez V. J., Arnalte-Mur P., Pons-Bordería M.-J., Pareja-Flores C., Paredes S., 2017, A&A, 601, A40 (HG17)
Itoh M., Inagaki S., Saslaw W. C., 1988, ApJ, 331, 45 (IIS88)
Itoh M., Inagaki S., Saslaw W. C., 1993, ApJ, 403, 476
Jones E. et al., 2020, Nature Methods, 17, 261
Kim J., Changbom P., L'Huillier B., Hong S. E., 2015, J. Korean Astron. Soc., 48, 213
Leicht O., Uhlemann C., Villaescusa-Navarro F., Codis S., Hernquist L., Genel S., 2019, MNRAS, 484, 269
McBride C. et al., 2011, Am. Astron. Soc. Meeting, 217, 249
Marinacci F. et al., 2018, MNRAS, 480, 5113
Martizzi D. et al., 2019, MNRAS, 486, 3766
Martizzi D., Vogelsberger M., Torrey P., Pillepich A., Hansen S. H., Marinacci F., Hernquist L., 2020, MNRAS, 491, 5747
Montero-Dorta A. D. et al., 2020, MNRAS, 496, 1182
Naiman J. P. et al., 2018, MNRAS, 477, 1206
Nelson D. et al., 2018, MNRAS, 475, 624
Nelson D. et al., 2019, Comput. Astrophys. Cosmol., 6, 2
Peebles P. J. E., 1980, The Large-Scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Perlmutter S. et al., 1999, ApJ, 517, 565
Pillepich A. et al., 2018, MNRAS, 475, 648
Planck Collaboration XIII, 2016, A&A, 594, A13
Python, 2019, Python Language Reference, version 3.7.6. http://www.python.org
Riess A. G. et al., 1998, AJ, 116, 1009
Rodriguez-Gomez V. et al., 2019, MNRAS, 483, 4140
Salvador A. I. et al., 2019, MNRAS, 482, 1435
Saslaw W. C., 1985, Gravitational Physics of Stellar and Galactic Systems. Cambridge Univ. Press, Cambridge
Saslaw W. C., 1986, ApJ, 304, 11
Saslaw W. C., 2000, The Distribution of the Galaxies. Cambridge Univ. Press, Cambridge
Saslaw W. C., Fang F., 1996, ApJ, 460, 16
Saslaw W. C., Hamilton A. J. S., 1984, ApJ, 276, 13
Sheth R. K., Saslaw W. C., 1996, ApJ, 470, 78
Springel V., 2010, MNRAS, 401, 791
Springel V. et al., 2018, MNRAS, 475, 676
Szapudi I., 1998, ApJ, 497, 16
Uhlemann C., Codis S., Pichon C., Bernardeau F., Reimberg P., 2016, MNRAS, 460, 1529
Uhlemann C. et al., 2018, MNRAS, 473, 5098
Weinberg D. H., Davé R., Katz N., Hernquist L., 2004, ApJ, 601, 1
Wen D., Kemball A. J., Saslaw W. C., 2020, ApJ, 890, 160
Yang A., Saslaw W. C., 2011, ApJ, 729, 123 (YS11)
York D. G. et al., 2000, AJ, 120, 1579

## APPENDIX A: ANALYSIS OF RESIDUES

In this section, we present a brief analysis of the residues in the CiC fits to the GQED and NBD models for TNG300-3 runs. The residues are shown in Fig. A1 as a function of redshift and cell size. As the counting cells are allowed to intersect, they form a statistical ensemble that is not entirely independent. Even if the cells were adjacent, objects belonging to nearby cells would be correlated. Hence, given the long range nature of the gravitational clustering, all cells are correlated in different degrees. In any case, the higher the number of cells in the ensemble used for the CiC computation, the better the statistics will be concerning the resulting form of $f_V(N)$. For the analysis of the goodness of the fitting models, we did not evaluate the $\chi^2$ estimates due to cell correlations; we use residual two norms of the fits, namely:

$$\mathrm{Res} = \sum_{N_0}^{N_{\max}} [f_V(N)_{\mathrm{sim}} - f_V(N)_{\mathrm{theo}}]^2 \,, \tag{A1}$$

where $N_0 \equiv (N = 0)$ and $N_{\max}$ is the largest number of galaxies in a cell.

We found that the residues (for both GQED and NBD models) were larger for smaller comoving cell sizes, especially for the smallest size. Except for the smallest cell, residues tended to increase at higher redshifts. The TNG300-3-H case, in which twice of initial counting cells was used, showed somewhat smaller residues (relative to each cell size) than the TNG300-3 runs at $z = 0$. For brevity, we omitted similar figures for the TNG100-3 runs, which showed similar trends; we briefly note that the dark-only TNG100-3 runs showed a significantly larger residue than in its full counterpart for the smallest cell. Overall, both GQED and NBD models showed similar residues.

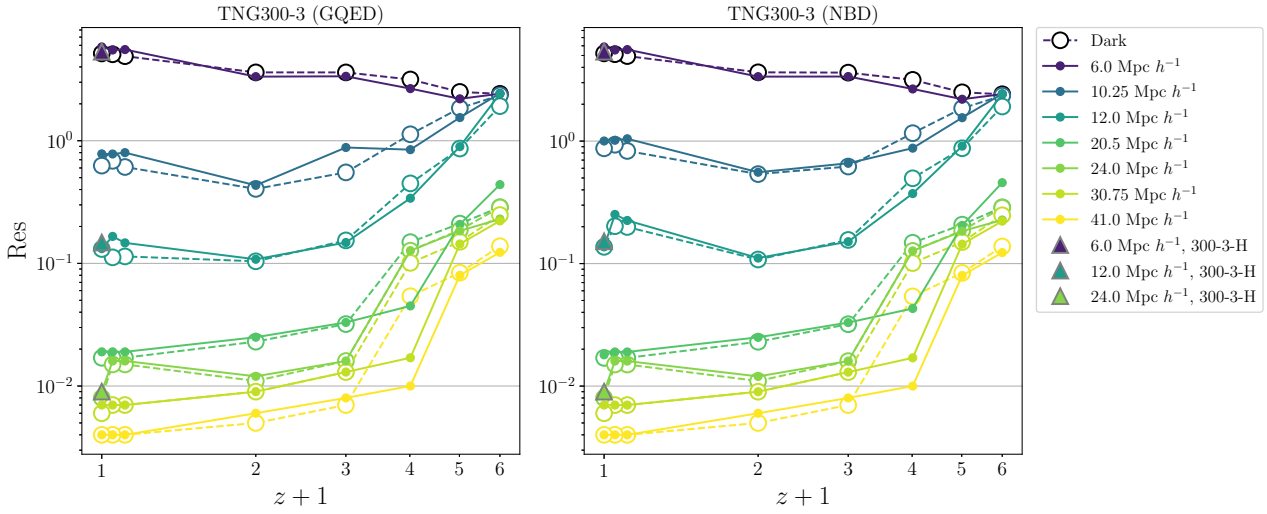


**Figure A1.** Residual two norms (in logarithmic scale) of the TNG300-3 CiC fittings as a function of redshift: GQED (left-hand panel) and NBD (right-hand panel).

This paper has been typeset from a TEX/LATEX file prepared by the author.