



**HAL**  
open science

## MoMaF : The Mock Map Facility

J. Blaizot, Y. Wadadekar, B. Guiderdoni, S. Colombi, E. Bertin, F. R.  
Bouchet, J. E. G. Devriendt, S. Hatton

► **To cite this version:**

J. Blaizot, Y. Wadadekar, B. Guiderdoni, S. Colombi, E. Bertin, et al.. MoMaF : The Mock Map Facility. Monthly Notices of the Royal Astronomical Society, 2005, 360, pp.159-17. 10.1111/j.1365-2966.2005.09019.x . hal-00009536

**HAL Id: hal-00009536**

**<https://hal.science/hal-00009536>**

Submitted on 13 Apr 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# MoMaF: the Mock Map Facility

Jérémy Blaizot,<sup>1,2,3\*</sup> Yogesh Wadadekar,<sup>1,4</sup> Bruno Guiderdoni,<sup>1</sup> Stéphane T. Colombi,<sup>1</sup> Emmanuel Bertin,<sup>1</sup> François R. Bouchet,<sup>1</sup> Julien E. G. Devriendt<sup>2,5</sup> and Steve Hatton<sup>1</sup>

<sup>1</sup>*Institut d'Astrophysique de Paris, 98 bis boulevard Arago, 75014 Paris, France*

<sup>2</sup>*Oxford University, NAPL, Keble Road, Oxford OX1 3RH*

<sup>3</sup>*LAM, Traverse du Siphon-Les trois Lucs, BP8-13376 Marseille, Cedex 12, France*

<sup>4</sup>*Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21210, USA*

<sup>5</sup>*CRAL, Observatoire de Lyon, 69561 St Genis Laval cedex, France*

Accepted 2005 March 15. Received 2005 March 3; in original form 2004 September 22

## ABSTRACT

We present the Mock Map Facility, a powerful tool for converting theoretical outputs of hierarchical galaxy formation models into catalogues of virtual observations. The general principle is straightforward: mock observing cones can be generated using semi-analytically post-processed snapshots of cosmological  $N$ -body simulations. These cones can then be projected to synthesize mock sky images. To this end, the paper describes in detail an efficient technique for creating such mock cones and images from the galaxies in cosmological simulations (GALICS) semi-analytic model, providing the reader with an accurate quantification of the artefacts it introduces at every step. We show that replication effects introduce a negative bias on the clustering signal – typically peaking at less than 10 per cent around the correlation length. We also thoroughly discuss how the clustering signal is affected by finite-volume effects, and show that it vanishes at scales larger than approximately one-tenth of the simulation box size. For the purpose of analysing our method, we show that number counts and redshift distributions obtained with GALICS/MOMAF compare well with  $K$ -band observations and the two-degree field galaxy redshift survey. Given finite-volume effects, we also show that the model can reproduce the automatic plate measuring machine angular correlation function. The MOMAF results discussed here are made publicly available to the astronomical community through a public data base. Moreover, a user-friendly Web interface (<http://galics.iap.fr>) allows any user to recover her/his own favourite galaxy samples through simple SQL queries. The flexibility of this tool should permit a variety of uses ranging from extensive comparisons between real observations and those predicted by hierarchical models of galaxy formation, to the preparation of observing strategies for deep surveys and tests of data processing pipelines.

**Key words:** methods: numerical – astronomical data bases: miscellaneous – galaxies: statistics – large-scale structure of Universe.

## 1 INTRODUCTION

Large galaxy surveys in which homogeneous data sets are acquired and analysed stand to benefit from the availability of mock images of galaxy wide/deep fields. These images are very useful for designing observational strategies and for putting predictions of various models into a format that can be directly compared with actual observations. Such an activity of ‘sky simulation’ is now running extensively as part of the data processing centres for satellite missions and for the next generation of ground-based instruments, because it is now acknowledged that data processing and mock observations

have to be integrated into instrument building from the beginning of the project to the interpretation of the data. However, such sky simulations are not easy, because they have to meet a certain level of realism to be useful. The purpose of this paper is to demonstrate how realistic galaxy catalogues and mock images can be synthesized from the outputs of a model of hierarchical galaxy formation.

Of course, mock galaxy surveys can be generated simply by drawing galaxy types, luminosities and sizes from their distribution functions. However, such an approach does not meet the requirements, because: (i) evolution cannot be included except in a very crude way, especially if we are to mimic number evolution; (ii) multi-wavelength surveys cannot be addressed easily; and (iii) spatial information cannot be addressed since galaxy positions are only known with a Poissonian distribution. Another approach is to

\*E-mail: jeremy.blaizot@oamp.fr

use observed fields that had been deliberately rescaled (Bouwens, Broadhurst & Silk 1998). A third approach is to start from theoretical priors and use numerical simulations to describe hierarchical clustering. In this approach, the main issue is how to transform mass to light and implement a robust method to obtain galaxies within dark matter (DM) structures.

There are basically two paths that this theoretical approach can take: (i) biasing schemes and (ii) halo models. The simplest approach (i) consists of ‘painting’ galaxies on dark matter simulations using phenomenological prescriptions to pick DM particles to be galaxies. Such methods, based on the linear bias formalism use the smoothed density field only (Cole et al. 1998) and are thus very efficient for big, low-resolution simulations (e.g. the Hubble volume simulations described by Evrard et al. 2002). The more subtle approach (ii) uses haloes identified in DM simulations. Several variations exist. In the simplest implementation, haloes are populated with galaxies according to a given halo occupation distribution (Peacock & Smith 2000; Scoccimarro et al. 2001; Berlind & Weinberg 2002) which may depend on luminosity (Yang et al. 2004). In a more sophisticated approach, a Monte Carlo scheme is used to build a merger history tree for each halo identified in the simulation, and a semi-analytic model (SAM) is used to evolve galaxies in these trees (e.g. Kauffmann, Nusser & Steinmetz 1997; Benson et al. 2000). In this approach, the SAM is used to generate a physically motivated halo occupation distribution. Eventually, a hybrid approach can be used, which extracts halo merging history trees from the dark matter simulations, and uses a SAM to evolve galaxies in these (Kauffmann et al. 1999; Hatton et al. 2003, hereafter GALICS I; Helly et al. 2003). Mock catalogues made using this technique have been made to mimic the CfA redshift survey (Diaferio et al. 1999) or the DEEP2 survey (Coil, Davis & Szapudi 2001). As a matter of fact, all the implementations of the two above methods are closely linked because the bias formalism often relies on analysis of the SAMs themselves (e.g. Cole et al. 1998; Somerville et al. 2001). They must therefore be considered as complementary rather than competing. The spirit of the physically motivated semi-analytic recipes can be extended to other objects such as X-ray clusters (Evrard et al. 2002). However, these approaches have generally been designed to fulfil the needs of a specific survey (for instance, 2dF for Cole et al. 1998 or DEEP2 for Coil et al. 2001). This limited scope can be extended in at least three ways.

First, it would be interesting to elaborate a generic approach to address the construction of mock observing cones from the outputs of  $N$ -body simulations at various cosmic times. The main issue is that, depending on the depth and solid angle of the observing cone, the finite size of the box may call for box replication along the line of sight (hereafter radial replication) and box replication perpendicular to the line of sight, at the same cosmic time (hereafter transverse replication). Wide-field shallow surveys, with negligible evolution, can be constructed mainly from a single box (the last output of a simulation corresponding to  $t = t_0$ ). In contrast, deep, pencil-beam surveys generally have to use numerous radial replications, whereas they may avoid transverse replication. Several issues have to be addressed here: the effects that replication might have on mock catalogues, the effect that using a finite volume might have on catalogues, and the sensitivity of catalogues to the number of time outputs of the root simulation. Of course, using a larger box size would improve the situation, but given finite computer resources (CPU time and memory), using a larger box would require a trade-off in the mass resolution of the simulation, which is not acceptable if galaxies are to be modelled with a sufficient level of realism. A Hubble volume would be the ideal situation avoiding any radial or

transverse replication, but so far, the largest volume simulation ( $\Lambda$  cold dark matter;  $\Lambda$ CDM with  $3000 h^{-1}$  Mpc on a side, and  $10^9$  particles) has a particle mass of  $m_p = 2.25 \times 10^{12} h^{-1} M_\odot$  (Evrard et al. 2002), much too large to address galaxy formation with any of the ‘hybrid models’. For instance, the  $\Lambda$ CDM simulation used in GALICS I has only  $100 h^{-1}$  Mpc on a side and  $m_p = 5.51 \times 10^9 h^{-1} M_\odot$ ; yet resolution effects are visible for galaxies fainter than  $L^*/8$  at  $z = 0$ . While we await a three orders of magnitude improvement of the simulations, addressing the replication issues is unavoidable if one wants to mimic large-volume observations with high resolution.

Secondly, the mock catalogues are useful if they gather together a large number of potentially observable properties. For instance, it is obvious that a mock catalogue designed to prepare and analyse a redshift survey of a magnitude-limited sample in a given photometric band, will incorporate at least the predicted redshifts and apparent magnitudes in that band. However, the redshift survey will also be used for follow-up at other wavelengths, and other studies (for instance, spectral classification once the spectra are properly calibrated). A good mock catalogue will be able to provide all of these pieces of information at wavelength bands different from those of the original survey. Ultimately the mock catalogue will enable the production of field images at many wavelengths, making source extraction using the same data processing pipeline as the actual observations possible.

Thirdly, the catalogues quickly become very large, and the question of accessibility to relevant information becomes crucial. Generally they are made available on Web pages as ASCII tables, mostly as galaxy catalogues from snapshots and more rarely as galaxy catalogues from observing cones. The more realistic these tables try to be (by including many galaxies with many properties), the more difficult to read and use they become, because of their growing size. The solution to this problem is to make the catalogues accessible through a data base that can be queried to make ad hoc subsamples fitting specific needs within a wide range of possibilities.

The purpose of this paper is to contribute along these three lines, by (i) presenting a package called MOMAF (i.e. Mock Map Facility) that generates observing cones from the outputs of our galaxies in cosmological simulations (GALICS) model and (ii) discussing in detail the limitations of the method. From these observing cones, synthetic catalogues are generated, that can be easily related to the catalogues of galaxies in the snapshots. The catalogues gather together a large number of properties, including magnitudes in many photometric bands of interest. These GALICS/MOMAF catalogues are made available in an on-line data base that can be queried through a simple Web interface at <http://galics.iap.fr>.

To illustrate our technique, we use examples drawn from the  $\Lambda$ CDM simulation and the GALICS post-processing described in GALICS I. This simulation is a compromise in terms of mass resolution and volume size, and gives a satisfactory description of the luminosity functions over typically 5 mag. However, the techniques we describe are generic, and can be used for larger simulations. In this study, we do not address in detail the drawbacks of our model (see GALICS I) in terms of mass resolution or limited volume, nor the quality of its predictions. We are only interested in how these predictions can be converted into mock observations. First examples of using these predictions and the mock catalogues can be found in Devriendt et al. (in preparation, hereafter GALICS II) and Blaizot et al. (2004, hereafter GALICS III). Two other papers will address issues which are more relevant to mock images: multi-wavelength faint galaxy counts (GALICS IV) and correlation functions (GALICS V).

This paper is organized as follows. Section 2 summarizes the main features of the GALICS model that are relevant to our study. In

Section 3, we describe our technique of catalogue and map building from the simulation snapshots. In Section 4, we explore the different limitations of our method, most of which are actually general enough to apply to other mock catalogues in the literature based on the tiling method. We explain in Section 5 how all the products of GALICS and MOMaF are stored in a relational data base accessible from the web, and illustrate a few key features of this data base. Section 6 contains a discussion concerning how these mock catalogues and images may be used, and presents perspectives for further developments.

## 2 THE GALICS MODEL

GALICS is a model of hierarchical galaxy formation which combines high-resolution cosmological simulations to describe the dark matter content of the Universe with semi-analytic prescriptions to deal with the baryonic matter. This hybrid approach is fully described in GALICS I and GALICS II and we only briefly recall its relevant features here.

### 2.1 Dark matter simulation

The cosmological  $N$ -body simulation we refer to throughout this paper was performed using the parallel tree-code developed by Ninin (1999). It is a flat cold dark matter model with a cosmological constant ( $\Omega_m = 0.333$ ,  $\Omega_\Lambda = 0.667$ ). The simulated volume is a cube of side  $L_{\text{box}} = 100 h_{100}^{-1}$  Mpc, with  $h_{100} = 0.667$ , containing  $256^3$  particles of mass  $8.272 \times 10^9 M_\odot$ , with a smoothing length of 29.29 kpc. The power spectrum was set in agreement with the present-day abundance of rich clusters ( $\sigma_8 = 0.88$ , from Eke, Cole & Frenk 1996), and we followed the DM density field from  $z = 35.59$  to 0, outputting 100 snapshots spaced logarithmically in the expansion factor.

In each snapshot we use a friend-of-friend algorithm to identify virialized groups of more than 20 particles, thus setting the minimum dark matter halo mass to  $1.65 \times 10^{11} M_\odot$ . We compute a set of properties for these haloes, including the position and the velocity of the centre of mass, kinetic and potential energies, and the spin parameter. Then, assuming a density profile for the virialized dark matter, we compute the virial radius a spherical halo would have to have the same mass and potential energy, thus making the link to the idealized semi-analytic approach.

Once all the haloes are identified in each snapshot, we compute their merging history trees, following the constituent particles from one output to the next. The merging histories we obtain are more complex by far than in semi-analytic approaches as they include evaporation of haloes, fragmentation, and several artefacts due to loose friend-of-friend identifications. The way we deal with these is described in detail in GALICS I.

### 2.2 Baryonic prescriptions, or how mass turns into light

When a halo is first identified, it is assigned a mass of hot gas, assuming a universal baryonic to dark matter mass ratio ( $\Omega_b = 0.045$  in our fiducial model). This hot gas is assumed to be shock heated to the virial temperature of the halo and in hydrostatic equilibrium within the dark matter potential well. The comparison of the cooling time of this gas to its free-fall time, as a function of the radius, yields the mass of gas that can cool to a central disc during a time-step. The size of this exponential disc is given by conservation of specific angular momentum during the gas in-fall and scales as the spin parameter of the halo. Then, the cooled gas is transformed into stars with a rate proportional to its mass divided by the disc dynamical time, with a given efficiency. The stars formed are distributed in mass according

to an initial mass function (IMF) taken from Kennicutt (1983). The stellar population of each galaxy is then evolved between the time-steps, using a substepping of at most 1 Myr. During each substep, stars release gas and metals in the interstellar medium (ISM), and we follow this gas recycling in time, assuming instantaneous mixing. The massive end of the stellar population soon explodes into supernovae which also release metals and energy into the ISM or the intergalactic medium (IGM). We model this as a function of the instantaneous star formation rate.

When two haloes merge, the galaxies they contain are gathered within the same final halo and their orbits perturbed. Subsequently, due to dynamical friction or satellite–satellite collisions, they can possibly merge. A ‘new’ galaxy is then formed (the descendant of the two progenitors) and the stars and gas of the progenitors are distributed in three components: a disc, a bulge and a starburst, with the amount of what goes where being fixed by the ratio of the masses of the two progenitors. The new galaxy can be elliptical (in shape) if the two progenitor galaxies have approximately the same mass, or remain a spiral if one of the merged galaxies has negligible mass.

The spectral energy distributions (SEDs) of our modelled galaxies are computed by summing the contribution of all the stars they contain, according to their age and metallicity, both of which we keep track of all along the simulation. Then, extinction is computed assuming a random inclination for disc components and the emission of dust is added to the extinguished stellar spectra with STARDUST (Devriendt, Guiderdoni & Sadat 1999). Finally, a mean correction for absorption through the intergalactic medium is implemented following Madau (1995), before we convolve the SEDs with the desired filters in the observer frame.

### 2.3 Resolution effects

The mass resolution of the DM simulations affect both the physical and statistical properties of modelled galaxies in the following ways.

(i) The particle mass of the cosmological simulation sets a minimum halo mass. Converting this halo mass into a galaxy mass, assuming that all the gas in the halo cools, one obtains a threshold mass above which our sample of galaxies is complete: the formal completeness limit. Below this mass, although we do have galaxies, our sample is not complete since we miss galaxies in undetected haloes. This direct effect of mass resolution is responsible for the lack of dwarf galaxies in the standard GALICS model. To express the completeness limit in terms of magnitudes is not straightforward because of the complex processes that convert mass into light. One can define a limiting magnitude, at a given redshift, such that, say, 95 per cent of the galaxies brighter than that will be more massive than the formal mass resolution. Because there is no one-to-one relation between mass and luminosity, however, the luminosity selection is in practice more drastic than the selection on mass. As an example, these magnitudes are given in several wavebands at  $z = 0$  and 3 in Table 1. They can easily be derived for other wavebands or redshifts from the GALICS data base (see Section 5).

(ii) In a Universe dominated by cold dark matter, small structures form first, and then merge and accrete material so as to evolve into larger haloes. In other words, the characteristic mass  $M_*$  of the mass distribution of haloes increases as redshift decreases. The mass resolution of our numerical simulation is fixed, however, and does not allow us to identify objects less massive than  $\sim 1.6 \times 10^{11} M_\odot$ , at any redshift. Hence, going back in time, more and more haloes are not resolved, and one eventually reaches a point where no halo can be detected. We call  $z_{\text{lim}}$  the redshift limit when this happens.

**Table 1.** 95 and 75 per cent completeness limits in terms of absolute rest-frame magnitudes at redshifts of 0 and 3. At  $z = 0$ , the magnitudes are expressed in the Vega system, and the filters are Johnson's. At  $z = 3$ , the magnitudes are expressed in the  $AB$  system, and the filters are those from Steidel & Hamilton (1993).

	Waveband	95 per cent completeness	75 per cent completeness
$z = 0$	<i>U</i>	−19.6 mag	−18.2 mag
	<i>B</i>	−19.5 mag	−18.5 mag
	<i>V</i>	−20.1 mag	−19.4 mag
	<i>R</i>	−20.6 mag	−20.1 mag
	<i>I</i>	−21.1 mag	−20.7 mag
	<i>J</i>	−21.8 mag	−21.5 mag
	<i>K</i>	−22.8 mag	−22.5 mag
$z = 3$	$U_n$	−20.6 mag	−20.3 mag
	<i>G</i>	−21.1 mag	−20.9 mag
	<i>R</i>	−21.1 mag	−20.9 mag

At higher redshifts, we miss all possible galaxies. In our simulation, one finds  $z_{\text{lim}} \sim 7$ .

(iii) A more subtle effect of resolution is that missing small structures means missing part of the histories of the galaxies. In practice, we showed that for our standard simulation, a galaxy needs to have evolved for  $\sim 1$  Gyr before its properties have converged (see GALICS III). Although this is virtually no constraint at  $z = 0$ , where most galaxies are much older than 1 Gyr, the constraint becomes drastic at  $z = 3$ , when the age of the Universe is only  $\sim 2$  Gyr. To ease the selection of mature galaxies for users of the data base (see Section 5), we assign the morphological type ‘Im’ to immature galaxies.

### 3 MOCK OBSERVATIONS

In this section, we explain how we convert the outputs of GALICS described above into mock observations. We first describe the inputs we need from GALICS or any other model/simulation of galaxy formation. Then, we show how these inputs are turned into mock maps, and point out the main limitation of our technique: replication effects. Finally, we briefly explain how we can project catalogues on to realistic pre- or post-observing maps.

#### 3.1 Inputs

The method we developed to generate mock catalogues from outputs (or snapshots) of cosmological simulations at a finite number of redshifts is general and can be used for a variety of objects (e.g. clusters or quasars). Here, we describe the features needed in these snapshots for galaxies.

The snapshots have to be (cubic) volumes of equal comoving size (in our standard simulation,  $L_{\text{box}} = 100 h^{-1}$  Mpc) with periodic boundary conditions. These snapshots must each contain the following information.

- (i) The redshift or expansion factor of the snapshot.
- (ii) The position of each galaxy within the snapshot.
- (iii) The velocity of each galaxy within the snapshot.
- (iv) The characteristic scalelength of each component of each galaxy (disc, bulge and burst).
- (v) The inclination of each galaxy, which was used to compute its extinction.

(vi) The absolute  $AB$  magnitude of each galaxy in the desired filters, computed in the observer frame as

$$M_{v_0}(z) = -2.5 \log \left[ \frac{L_{v_0(1+z)}}{(10 \text{ pc})^2} \right] - 2.5 \log(1+z) + 48.6, \quad (1)$$

where  $z$  is the redshift of the snapshot, and

$$L_{v_0(1+z)} = \int (1+z) f_{v_0}(v) L[v(1+z)] dv \quad (2)$$

is the luminosity of a galaxy at redshift  $z$ , through a normalized filter response  $f_{v_0}$ . Note that because the peculiar velocities or positions relative to the observer are not known at this stage, we use the redshift  $z$  of the snapshot to compute these magnitudes. This approximation will be corrected for when we compute apparent magnitudes (see Section 3.2.2). Also note that these magnitudes take into account extinction by the intergalactic medium, computed at the redshift of the snapshot.

(vii) The first-order derivatives of the above magnitudes with redshift, in each filter. For galaxies in snapshot  $i$ , these derivatives are estimated as

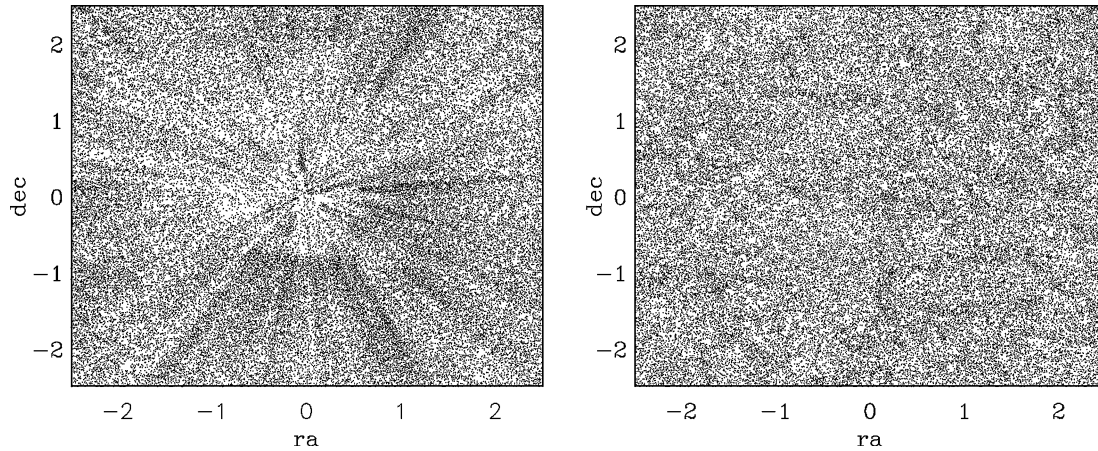
$$\frac{dM}{dz} = \frac{M[z(i-1)] - M[z(i)]}{z(i-1) - z(i)}, \quad (3)$$

where  $z(i)$  is the redshift of snapshot  $i$  [in our convention,  $z(i) < z(i-1)$ ] and  $M[z]$  is the observer-frame absolute magnitude assuming the galaxy is at redshift  $z$  (equation 1). Note that this expression does not account for the evolution of galaxies, as it involves the magnitudes of the same galaxy put at different redshifts. Equation (3), however, captures the  $K$ -correction and variations of IGM extinction with  $z$ , which are the main drivers of average variations of apparent properties with redshift in mock catalogues (see Section 3.2.2 and 4.3.1).

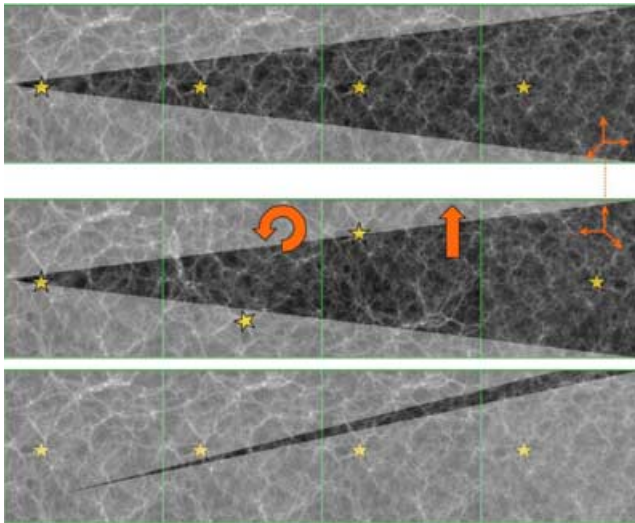
All of the above quantities are direct outputs of GALICS, except for positions and velocities. These are computed as a post-treatment, using information from the DM simulations (positions and velocities of the haloes) and from GALICS (orbital radii). The position of a galaxy within a snapshot is thus defined by  $\mathbf{g} = \mathbf{h} + r_{\text{orb}} \times \mathbf{u}$ , where  $\mathbf{h}$  is the position of its host halo,  $r_{\text{orb}}$  is the orbital radius of the galaxy and  $\mathbf{u}$  is a normalized vector of random direction. The peculiar velocity of a galaxy is defined as  $\mathbf{v}_g = \mathbf{v}_h + \delta\mathbf{v}$ , where  $\mathbf{v}_h$  is the peculiar velocity of its host halo and  $\delta\mathbf{v}$  is the peculiar velocity of the galaxy within this halo. The amplitude of  $\delta\mathbf{v}$  is drawn randomly from a Gaussian distribution of width equal to the circular velocity of the halo, and its direction is random. Note that the velocities of central galaxies are taken to be that of the centre of mass of their host haloes.

#### 3.2 Mock catalogues

Such inputs, corresponding to the same simulated region of universe at different redshifts, will cause replication effects when piled in a mock lightcone, namely the regular repetition of structures in mock catalogues or images. Transverse replications are due to the fact that the same volume, in the same state of evolution, is used several times to fill an observing cone across the line of sight. Radial replications occur because the same volume, although taken at different cosmic times, is repeatedly used to fill the observing cone along the line of sight. Because the largest structures evolve slowly (i.e. over several time-steps), they will create pseudo-periodicity in mock catalogues or mock maps. In Fig. 1 (see also Fig. 2), we show how replications create an artificial perspective effect in catalogues (left-hand-side panel). Replication effects can be suppressed with the ‘random tiling’ method (right-hand-side panel), which we describe here.



**Figure 1.** Angular projection of mock catalogues of  $5 \times 5 \text{ deg}^2$  (the ‘r.a.’ and ‘dec.’ labels are arbitrary labels for two orthogonal directions on the mock sky). Each point represents a bright galaxy. The left-hand-side panel shows a catalogue in which the snapshots were piled together without any randomization. This leads to replication effects, similar to a perspective effect (in an expanding/evolving universe). The right-hand-side panel shows a catalogue containing the same galaxies, which was made using the random tiling method. All replication effects have disappeared.



**Figure 2.** Illustration of the cone-making process. In the top panel we show the result of a straightforward tiling of snapshots. In this case, the structures (such as the star) appear repeatedly along the line of sight. The middle panel shows the effect of the three types of transformation we apply randomly to each underlying box. Thanks to rotations, translations and inversions, the underlying boxes are decorrelated from one another. In the lower panel we show how it is possible to avoid reshuffling underlying boxes to generate a pencil-beam type field.

### 3.2.1 Random tiling

Building a catalogue from the inputs described above consists of distributing the simulated galaxies in an observing cone, and computing their apparent properties in this new geometry. First, we define a three-dimensional (3D) pavement of cubic underlying boxes of side  $L_{\text{box}}$  ( $=100 \text{ comoving } h^{-1} \text{ Mpc}$ ). Then, we fill the underlying boxes inside the lightcone with galaxies in the following way:

(i) determine the time-steps  $i = n, \dots, n + k$ , which will be needed in order to fill the current underlying box, knowing that time-step  $i$  will be used to fill the lightcone between  $[z(i-1) + z(i)]/2$  and  $[z(i) + z(i+1)]/2$ ;

(ii) to each of these snapshots, apply the same transformation, which is a random combination of the following transformations:

- (a) a shift of random amplitude (between 0 and  $L_{\text{box}}$ ) in each of the three directions ( $x, y, z$ );
- (b) a rotation of  $0, \pi/2, \pi$  or  $3\pi/2$  around each axis;
- (c) the inversion of one of the axes picked randomly (e.g.  $x \mapsto -x$ ), or none;

(iii) use the transformed positions and velocities of galaxies to include them in the lightcone and compute their apparent properties;

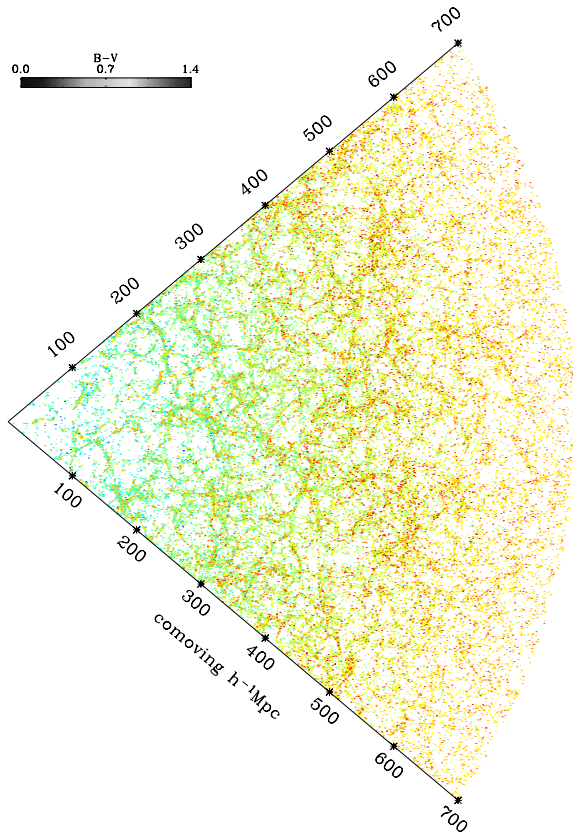
(iv) move on to the next underlying box, and repeat the previous steps until the lightcone is filled.

The first step allows a galaxy in the cone to be taken from the output box which has the closest redshift to the redshift relative of the galaxy to the observer. This has the advantage of picking galaxies at a stage of evolution as close as possible to that they would have if we had continuous outputs.

In the second step, the shifting, rotating and inverting of the underlying boxes is performed to suppress replication effects. The shuffling of the underlying boxes, outlined with thick lines in Fig. 2, decorrelates them from one another, thus suppressing replication effects and any information on scales larger than the box size. Although breaking the continuity of the density field makes us lose a fraction of the spatial information (see Section 4.1), we chose this solution because we have good control on this information loss.

For deep pencil-beam surveys it is possible to avoid replication effects simply by choosing an appropriate line of sight so that the lightcone will intersect different regions of each underlying box. It is better in this configuration not to shuffle the boxes so as to keep all the spatial information. This is an option which is implemented in our code, and illustrated on the lower panel of Fig. 2. Note, however, that, although the density field is continuous throughout the cone, clustering information on scales larger than  $L_{\text{box}}$  is still missing, because it is not contained in the DM simulations: replication effects can be suppressed but not finite-volume effects (see Section 4.2).

It is also possible to chose the position of the observer, relative to the first underlying box, in both options. This allows one to test for cosmic variance on local sources, and is useful for understanding the statistical significance of the bright end of galaxy counts.



**Figure 3.** Mock 2dFGRS: each point represents a galaxy brighter than  $b = 19.5$  from a lightcone of  $75 \times 4 \text{ deg}^2$ . The cone was truncated at a comoving distance of  $700 h^{-1} \text{ Mpc}$ , corresponding to  $z \sim 0.23$ . The shades of the points indicate the apparent  $B - V$  colours of the galaxies according to the shaded table in the upper left-hand corner.

### 3.2.2 Apparent magnitudes

Because we use a finite number of time-steps, (i) galaxies are picked at a cosmic time which is different from that corresponding to their distance in the mock lightcone, (ii) the SEDs are not convolved with the filters at the exact redshifts and (iii) IGM extinction is not computed for the correct redshifts. Point (i) means that an individual galaxy is not taken at the stage of evolution it would have in the case of continuous outputs. However, this does not affect the statistical properties of the mock catalogues because the overall galaxy population does not evolve much between time-steps, on average. This issue is also discussed in Section 4.3.1

Points (ii) and (iii), we correct for as follows. We define corrected observer-frame absolute magnitudes as

$$M_{\text{cor}} = M[z(i)] + \frac{dM}{dz}[z(d) - z(i)], \quad (4)$$

where  $M[z(i)]$  is the observer-frame absolute magnitude computed by GALICS at redshift  $z(i)$  of time-step  $i$  (equation 1),  $z(d)$  is the redshift of the galaxy evaluated from its comoving distance  $d$  to the observer in the mock lightcone and taking into account the peculiar velocity of the galaxy along the line of sight, and  $dM/dz$  is defined in equation (3). Note that this derivative only accounts for distance effects ( $K$ -correction and IGM extinction) and not evolution (point i above). The apparent magnitude of a galaxy is then obtained with

the luminosity distance  $d_L$ :

$$m = M_{\text{cor}} + 5 \log \left( \frac{d_L}{10 \text{ pc}} \right). \quad (5)$$

Thanks to the first-order correction of magnitudes, the distribution of galaxies in apparent colour–colour plots is continuous. This is especially important for colour selections of distant galaxies as shown in Blaizot et al. (2004).

In Fig. 3, we show an example lightcone with a detection limit close to that of the two-degree field galaxy redshift survey (2dFGRS; Colless et al. 2001). Each point represents a galaxy with  $b < 19.5$ , and the colours indicate the apparent  $B - V$  colour of the galaxies.

### 3.3 Mock maps

Two types of maps are useful to address different issues.

(i) Pre-observation maps are a simple projection of a mock catalogue on the sky. The only additional assumption required here is the functional form for the galaxy light profiles (e.g. an exponential disc).

(ii) Post-observation maps include, in addition, realistic modelling of the characteristics of the telescope–detector combination (e.g. diffraction effects, readout noise, photon shot noise). Where appropriate, atmospheric effects can also be included (e.g. seeing, air glow). SKYKAKER (Erben et al. 2001) is a useful tool for producing post-observation maps.

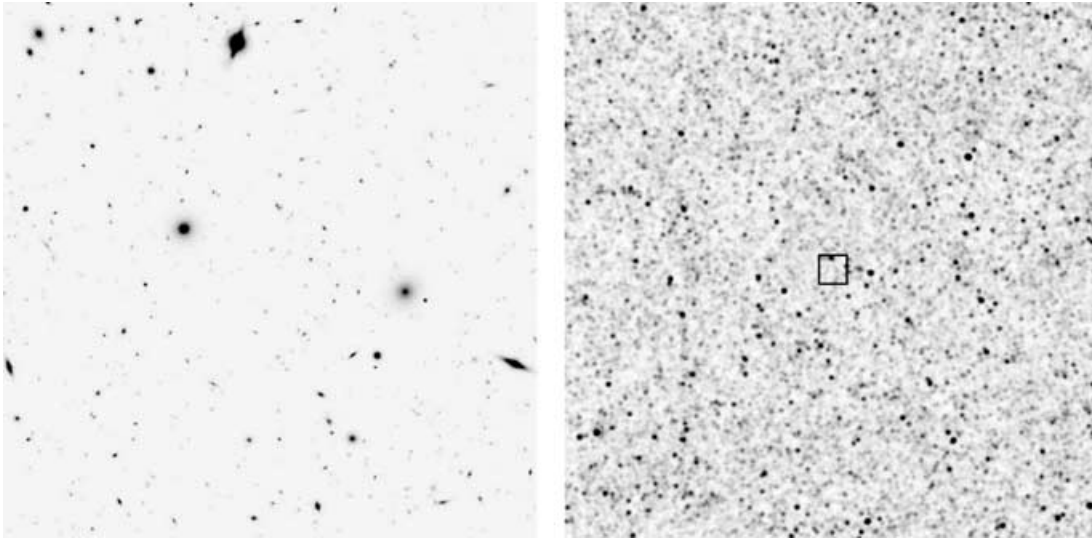
#### 3.3.1 Pre-observation maps

Consistent with the modelling of galaxies in GALICS, we display disc components with an exponential profile, and bulges and starbursts with a Hernquist profile (Hernquist 1990, equations 32–34). The profiles are truncated at approximately 10 times the half-mass radius of the component, and the Hernquist profile is also dimmed exponentially starting at five times the half-mass radius. To gain speed, we build a series of face-on disc and bulge templates on grids of different resolutions (e.g. from  $2 \times 2$  to  $2048 \times 2048$  pixel), each one normalized to unity. When adding the contribution of a galaxy to the final map, we chose the template which has a resolution just above that of the final map. For bulges, as they are assumed to be spherically symmetric, the projection is straightforward, and only rescaling of the template to the size of the component is required. For discs, we first have to flatten the template to account for inclination,<sup>1</sup> and then to rotate it to map the orientation of the disc. Eventually, we project the transformed templates on the final map grid, multiplying each template by the flux of the component it represents. The total flux on the final map is thus the sum of the fluxes of all galaxies in the lightcone, except when they are truncated on the border of the image.

Some aliasing effects appear because of the projection of the tilted template grids on the final map, but these will be washed out when the map is convolved with a PSF afterwards. Since the aim of this tool is to produce pre-observation maps (with a resolution that should be higher than the final post-observation map), there is not much point in correcting this effect via bilinear interpolation or other CPU-expensive methods.

Note that the images produced in this way are not limited in magnitude (up to the resolution limit of the simulation) and include

<sup>1</sup> This inclination is the same as that used in GALICS to compute the extinction of light by dust.



**Figure 4.** Left-hand-side panel: mock map of a  $3 \times 3$  arcmin<sup>2</sup> field, in the  $R$  band. Right-hand-side panel: mock map of a  $1 \text{ deg}^2$  field, at  $170 \mu\text{m}$ , convolved with a Gaussian PSF of width 10 arcsec. The central square in this image outlines the field shown in the optical range in the left-hand-side panel.

the contributions of all galaxies in the cone. It is important that all sources are added (even though some may be fainter than the detection limit) for estimating the background intensity. This is particularly relevant to far infrared or submillimetre surveys which are limited by confusion and where the background contains a significant part of the information.

Example mock maps are shown in Fig. 4. The left-hand-side panel shows an optical view ( $R$  band) of a  $3 \times 3$  arcmin<sup>2</sup> field, and the right-hand-side panel shows the far IR view of the same field (at  $170 \mu\text{m}$ ). This latter image was convolved with a Gaussian PSF of width 10 arcsec to mimic an observation by the PACS instrument on-board Herschel. No noise was added to these mock maps.

Finally, note that MOMAF allows one to generate all-sky maps, using the HEALPIX pixelization (Górski et al. 2002) chosen by the Planck consortium.

### 3.3.2 Post-observation maps

It is considerably more difficult to generate post-observing maps because separate modelling is required for each telescope–detector combination. MOMAF is designed to feed instrument numerical simulators (INS) with realistic catalogues or pre-observation maps. In the optical and near-infrared domain, a ready general tool for post-observing map generation is available in SKYKAKER (Erben et al. 2001). We briefly discuss this general INS here as an example of MOMAF possibilities.

SKYKAKER is an image simulation program, originally designed to assess SEXTRACTOR detection and measurement performance (Bertin & Arnouts 1996). The code (currently at version 2.3.4) has been much improved since. It is capable of simulating star and galaxy images with a high level of accuracy. Galaxies in SKYKAKER are modelled as a combination of a de Vaucouleurs bulge and an exponential disc. Various sources of noise and convolution with the point spread function can be included as desired.

There are two input files required for SKYKAKER to generate an image. One is a configuration file specifying the characteristics of the telescope and the detector and the seeing conditions. The second is the source list, which can include stars and galaxies. A typical line for a galaxy in a source list for SKYKAKER includes the ‘total’ magni-

tude, the bulge-to-total luminosity ratio, the bulge equivalent-radius in arcsec, the projected bulge aspect ratio, the bulge position angle in degrees, the disc scalelength in arcsec, the disc aspect ratio and the disc position angle in degrees. All of these are natural outputs of GALICS/MOMAF, as discussed in the previous section. There is no provision for adding starburst components in SKYKAKER. We use a workaround for such cases. We add the burst components from GALICS as additional bulges, with a scalelength obtained from GALICS, a bulge-to-total luminosity ratio of 1.0 and the appropriate starburst magnitude as the ‘total’ magnitude.

## 4 LIMITATIONS OF THE METHOD

Ideally, one should build a mock catalogue from a simulated volume much larger than the lightcone, and output the physics smoothly by propagating photons towards the observer through the expanding simulated universe. Although such simulations are becoming feasible today, their computational cost is still prohibitive. The method we propose with MOMAF stems from the same philosophy as GALICS and consists of extracting as much information as possible from a given simulation, and using that to build realistic catalogues at relatively low computational expense. Of course, however sophisticated the method we use, several limitations appear in MOMAF mock catalogues because they are built from the replication of finite information. The purpose of this section is to understand how the replication process affects our predictions.

The most important limitations of MOMAF result from the fact that we use a finite volume to describe the whole Universe. In order to do so, we have to replicate the simulated volume many times along and across the line of sight. Now, because we use the random tiling method to proceed with these replications, some clustering information is lost. This results in a negative random tiling bias which is discussed in Section 4.1. A more subtle effect comes from the fact that the finite volume of the simulation used to build a mock catalogue does not describe density fluctuations on large scales. Thus these fluctuations will be missing from the mock catalogues. This results in biases on count variance estimates and correlation function estimates. These finite-volume effects are described in Section 4.2.



In this section, we also check that other possible effects are under control. In Section 4.3.1, we investigate the effect of a finite time-step on the apparent properties of galaxies in mock catalogues. In Section 4.3.2, we check on the impact of mass resolution of the root simulation on different observable statistics.

#### 4.1 Random tiling bias

A negative bias on correlation functions is introduced in mock catalogues by the random tiling approach, which comes from the fact that we decorrelate pairs of galaxies from one underlying box to the other when reshuffling them to suppress (periodic) replication effects. Here, we first estimate this bias on the spatial two-point correlation function, and then project the results to derive the bias on the angular correlation function (ACF).

##### 4.1.1 Spatial correlation function

The spatial correlation function (SCF) can be computed by measuring the number of pairs of objects separated by a given distance. If one uses the estimator of Landy & Szalay (1993, hereafter LS93)

$$\xi(r) = \frac{DD(r) - 2DR(r) + RR(r)}{RR(r)}, \quad (6)$$

where  $DD(r)$ ,  $DR(r)$  and  $RR(r)$  are the number of data–data, data–random and random–random pairs separated by distance  $r$ , respectively, one obtains, through the logarithmic differentiation, the relative error on  $\xi(r)$ :

$$\frac{\delta\xi(r)}{\xi(r)} = \frac{\delta DD(r)}{DD(r)}, \quad (7)$$

because the random sample is not affected by the replication technique.

Now, take a galaxy near the edge of an underlying box (say at a distance  $d < r$  from the edge). The mean number of pairs that will be missed for this galaxy, due to replication, is

$$\delta\widetilde{DD}_g(d, r) = 2\pi r^2 n \xi(r) (1 - d/r) dr, \quad (8)$$

where  $n$  is the mean density of galaxies and the subscript  $g$  denotes that  $DD$  is the number of pairs lost for one galaxy. The tilde over  $DD$  denotes that we only consider pair loss through one side of the box. To compute the loss of pairs due to one side of the underlying box, we integrate the previous equation over  $d$ , from 0 to  $2r$ , namely

$$\begin{aligned} \delta\widetilde{DD}(r) &= 2\pi r^2 n \xi(r) dr \int_{d=0}^{d=2r} n L_{\text{box}}^2 (1 - d/r) dd \\ &= 2\pi r^3 L_{\text{box}}^2 n^2 \xi(r) dr. \end{aligned} \quad (9)$$

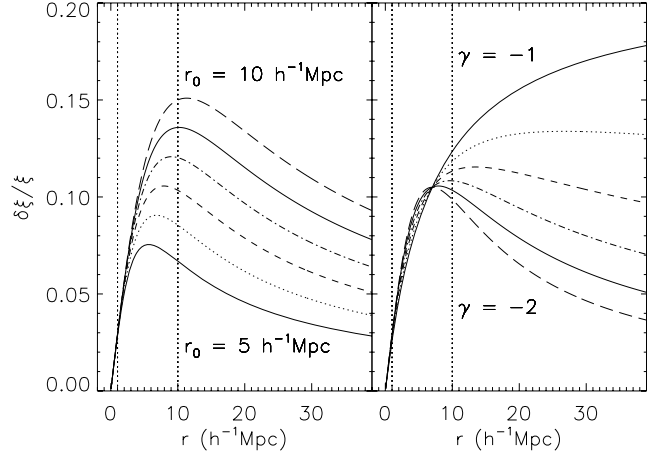
Note that we neglected corner effects here. This is justified by the fact that one should always consider separations much smaller than the size of the box (i.e.  $r \ll L_{\text{box}}$ ).

To obtain the pair loss over a whole box, simply multiply the previous result by 6 (the number of sides):

$$\delta DD(r) = 12\pi r^3 L_{\text{box}}^2 n^2 \xi(r) dr. \quad (10)$$

Had we not broken the continuity in the density field between each underlying box, the number of pairs would simply be, for a whole box:

$$DD(r) = 4\pi r^2 dr n [1 + \xi(r)] \times n L_{\text{box}}^3, \quad (11)$$



**Figure 5.** Expected relative underestimation on the spatial correlation function, assuming  $\xi(r) = (r/r_0)^\gamma$ . The left-hand-side panel shows variations with  $r_0$ , with  $r_0 = 5, 6, 7, 8, 9, 10 h^{-1}$  Mpc from the bottom curve to the top curve, with  $\gamma$  being fixed to  $-1.8$ . The right-hand-side panel shows the dependence on  $\gamma$ , with  $\gamma$  spanning the range  $[-2, -1]$ , from bottom to top, with  $r_0$  being fixed to  $7 h^{-1}$  Mpc. On each panel, the right-hand side vertical line shows the approximate upper limit of validity of measurements of  $\xi$  (one-tenth of  $L_{\text{box}}$ ). The left-hand-side vertical line roughly indicates the size of a cluster, below which our spatial information is uncertain.

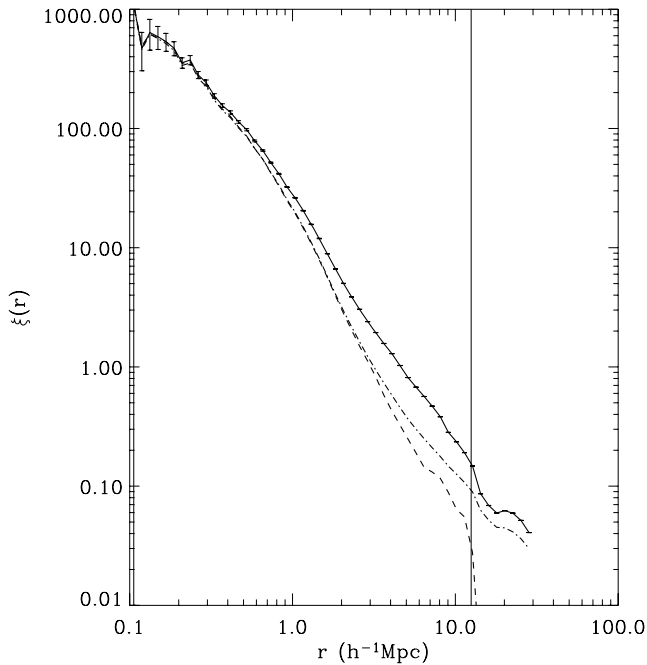
where the first right-hand term is the number of pairs expected for one galaxy and the second right-hand term is the number of pairs in the box. The relative error in the number of pairs, due to shifting the underlying boxes is thus

$$\frac{\delta DD(r)}{DD(r)} = 3 \frac{r}{L_{\text{box}}} \frac{\xi(r)}{1 + \xi(r)}. \quad (12)$$

For a numerical estimate, consider the Lyman break galaxies (LBG) population. For these galaxies, we expect  $\xi(r) = 1$  at  $r_0 \sim 6 h^{-1}$  Mpc. Thus, for our simulation, with  $L_{\text{box}} = 100 h^{-1}$  Mpc, one finds  $\delta\xi/\xi < 10$  per cent for LBGs, at  $6 h^{-1}$  Mpc.

In Fig. 5, we show the theoretical underestimation on spatial correlation function measurements from our catalogues. In the plots, we assume a correlation of the form  $\xi(r) = (r/r_0)^\gamma$ , and we let  $r_0$  and  $\gamma$  vary. For a wide range of these two parameters, the error due to transforming the underlying boxes is less than 10 per cent from 1 to  $10 h^{-1}$  Mpc.

In Fig. 6, we show a measure of the bias on  $\xi$  introduced by randomizing the boxes. To do this, we cut our snapshot volume into  $8^3$  subboxes to which we applied translations, rotations and inversions as described above. We then measured  $\xi(r)$  on the original snapshot (solid curve) and on the shuffled snapshot (dashed line). In Fig. 6 we plotted the prediction for the bias according to the above calculation as a dot-dashed line. The agreement between the measurement and the analytical prediction is very good on scales up to  $\sim 1/5 L_{\text{box}}$  (with  $L_{\text{box}}$  being  $100/8 h^{-1}$  Mpc, as shown by the vertical line in Fig. 6). In the previous section, we showed that this scale is where finite-volume effects come into play. Although most finite-volume effects are not present here because we use all the subboxes to fill the simulated volume, the randomization of subboxes kills any signal at scales larger than a subbox. This is not described by the above analytic calculation and results in the sharp cut-off of  $\xi$  at  $\sim 1/5 L_{\text{box}}$ .



**Figure 6.** Measured spatial correlation function of the dark matter particles in the  $z = 0$  snapshot (solid curve) and reshuffled on the same snapshot (dashed line). Also plotted is the prediction from equation (12) (dot-dashed line). The error bars attached to the solid line are Poissonian errors. The vertical line indicates the size of subboxes used for the reshuffling.

#### 4.1.2 Angular correlation function

We can use the bias on the SCF derived above to evaluate that on the ACF. Let us first remember that, in the small angle approximation, the angular correlation is related to the spatial one by (Bernardeau et al. 2002),

$$w(\theta) = \int d\chi \chi^4 D_M(\chi) \psi^2(\chi) \int_{-\infty}^{\infty} \theta \xi(r) dx, \quad (13)$$

where  $\chi$  is the radial distance,  $D_M$  is the angular distance ( $D_M = \chi$  in a flat universe),  $\psi$  is a selection function satisfying  $\int \chi^2 \psi(\chi) d\chi = 1$  and  $r$  is the separation distance, related to the angular separation  $\theta$  and the integration variable  $x$  through the relation  $r = D_M \theta (1 + x^2)^{1/2}$ . If we now introduce the bias on  $\xi$  as  $\xi \mapsto \xi + \delta\xi$ , with  $\delta\xi$  given by equation (12), we can derive the corresponding bias on the ACF from equation (13):

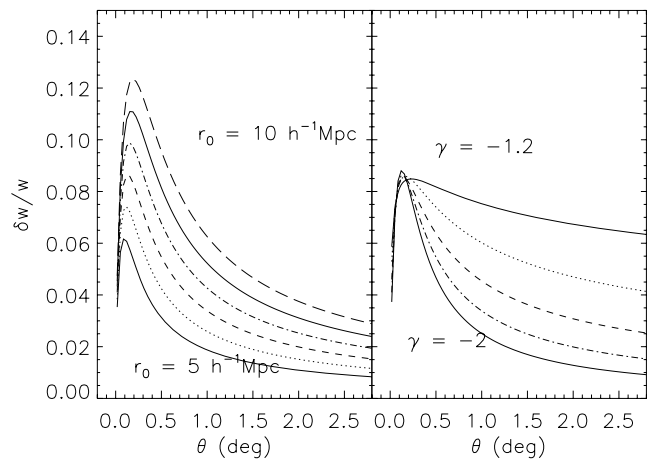
$$\delta w(\theta) = \int d\chi \chi^4 D_M(\chi) \psi^2(\chi) \int_{-\infty}^{\infty} 3\theta \frac{r}{L_{\text{box}}} \frac{\xi^2(r)}{1 + \xi(r)} dx. \quad (14)$$

Assuming that  $\xi$  can be written as the power law  $(r/r_0)^{-\gamma}$  and using  $D_M(\chi) = \chi$  in a flat universe, we end up with

$$\delta w(\theta) = \frac{3r_0^\gamma \theta^{2-\gamma}}{L_{\text{box}}} \int_0^\infty d\chi \chi^{6-\gamma} \psi^2(\chi) \times \int_{-\infty}^{\infty} \frac{(1+x^2)^{1/2-\gamma} dx}{(\chi\theta/r_0)^\gamma + (1+x^2)^{-\gamma/2}}. \quad (15)$$

Note that for this result, we also assumed that the SCF does not vary with redshift. This is obviously wrong in general but is justified if the selection function  $\psi$  is narrow enough (e.g. for LBGs).

Finally, using equations (13) and (15), and deciding on a selection function, one can compute numerically the relative bias induced on ACF measurements by the transformations of underlying boxes. An example is given in Fig. 7.



**Figure 7.** Expected relative underestimation on the angular correlation function, assuming  $\xi(r) = (r/r_0)^\gamma$ . The left-hand-side panel shows variations with  $r_0$ , with  $r_0 = 5, 6, 7, 8, 9, 10 h^{-1}$  Mpc from the bottom curve to the top curve, with  $\gamma$  being fixed to  $-1.8$ . The right-hand-side panel shows the dependence on  $\gamma$ , with  $\gamma$  spanning the range  $[-2, -1]$ , from bottom to top, with  $r_0$  being fixed to  $7 h^{-1}$  Mpc. The selection function chosen here is simply a top hat centred at  $\chi = 2000 h^{-1}$  Mpc and of width  $500 h^{-1}$  Mpc.

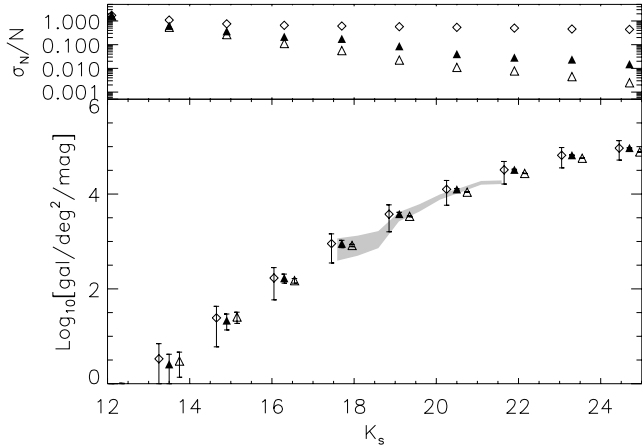
## 4.2 Finite-volume effects

Several limitations arise because we use a finite volume to describe the whole Universe. They are basically due to the fact that a finite volume  $V$  does not describe density fluctuations on scales typically larger than  $\sim V^{1/3}$ . In other words, although the mean number of galaxies in a simulation can be tuned to fit that observed in the Universe, the simulation does not describe the dispersion about this mean value. How this affects statistics from our catalogues is the question we address in this section. The simplest statistic we are interested in is galaxy counts, as a function of magnitude or redshift. Mock catalogues can be used in two ways: (i) to normalize models and (ii) to estimate errors (including cosmic variance). In Section 4.2.1, we discuss how finite volume affects both the counts and their variance. Then, in Section 4.2.2, we describe the bias on correlation functions introduced by finite-volume effects.

### 4.2.1 Effects on estimates of count variance

Two variances are relevant for counts in mock catalogues. The first is the variance which tells us about the dispersion of the number counts from mock catalogues each generated from a different simulated volume. The second is the variance that describes the dispersion in the number counts from mock catalogues made from a unique simulated volume. This variance tells us to what extent we can estimate cosmic variance with mock catalogues based on a given simulation. Let us proceed to virtual experiments to understand these quantities.

(i) Imagine we have a large number  $N$  of simulations at hand, all describing an equal volume  $V$ , but with initial conditions drawn from a much larger volume. From each simulation, we build a mock catalogue using MOMAF, and then count galaxies brighter than some magnitude limit. Finally, we measure the variance  $\sigma_1^2$  of the counts obtained in this way. Now, imagine that we also have  $N$  mock catalogues, each generated using an ideal technique and a simulated volume much larger than that of the lightcone. Call  $\sigma_2^2$  the variance in the counts measured from these catalogues. In the case where the volume of the cone is much smaller than volume  $V$ , the above two



**Figure 8.** Effect of finite volume on  $K$ -band counts. Lower panel, the filled (respectively, open) triangles show the mean number counts obtained from 20 catalogues made from our standard simulation (respectively, from one subbox), with the error bars giving the standard deviation about this mean. The open diamonds show the mean number counts measured from 125 mock catalogues, each built from a different subbox. The shaded area shows the locus of the K20 counts from Cimatti et al. (2002b). Upper panel, the relative standard deviation of the counts (same symbol code as in the lower panel).

variances will be equal. In the more realistic case where  $V$  is smaller than the volume of the cone, one will measure that  $\sigma_1 > \sigma_2$ : replication enhances the bias of the simulated volume, thus dispersing more counts from catalogues.

(ii) As a second experiment, imagine one has a unique simulated volume  $V$  as above, and builds many mock catalogues from it. These catalogues will be different from one another because of the random tiling process and because the lightcone may intersect different sections of  $V$  in different realizations. As before, measure the variance  $\sigma_3^2$  of the counts from these catalogues. Again, if the simulated volume is much larger than that of the lightcone, one will measure  $\sigma_3^2 \sim \sigma_2^2$ . In this case, one can use mock catalogues to estimate cosmic variance. However, in the case where the volume of the cone is larger than  $V$ ,  $\sigma_3$  will be found to be lower than  $\sigma_2$ , because replications do not add large-scale fluctuations. In the extreme case where the cone is very large compared with the simulated volume,  $\sigma_3$  will tend to zero, because the cone encloses all the information contained in  $V$ .

In Fig. 8, we show  $K$ -band counts measured from various mock catalogues having the geometry defined in the third line of Table 2 (‘counts’ catalogues). The shaded area shows the locus of K20 counts from Cimatti et al. (2002b), including Poissonian er-

ror bars. The filled symbols and their error bars give the mean and standard deviation for counts measured from 20 mock lightcones made using the standard simulated volume. Then we cut our root simulation into 125 subboxes, and made a mock catalogue out of each subbox. The open diamonds give the mean and standard deviation of counts measured from these 125 mock lightcones. Finally, the open triangles give the mean and standard deviation of counts measured from 20 mock catalogues made from a single subbox. The upper panel of Fig. 8 compares the relative standard deviations of these three measures.

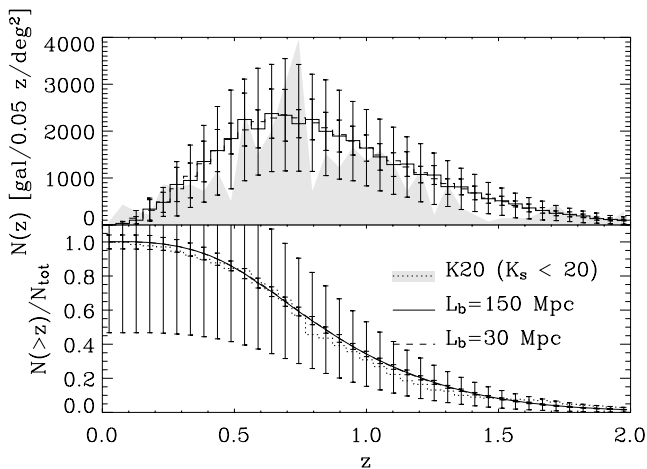
First, note that changing the size of the volume used to make mock catalogues does not change the shape of the counts, it only changes the amplitude. This was expected since evolution is the same in all subboxes. This tells us that if our root simulation is well normalized, counts from mock catalogues are not sensitive to the size of the simulated volume and can thus be used with confidence to normalize theory to observations.

Secondly, consider the difference between open and filled triangles in the upper panel of Fig. 8, i.e. the estimates of cosmic variance from cones made using one subbox or the full simulated volume. As expected, we find that using a small box leads to an underestimate of the variance. Table 2 tells us that the number of full boxes used to describe galaxies brighter than  $K_s = 20$  is approximately 12 along the line of sight (up to the median redshift) and 1/5 across the line of sight at the median redshift. This situation is at the limit where we can correctly estimate cosmic variance, since the lightcone only intersects a fraction of the box volume in each underlying box. For the subbox case, the lightcones include a full underlying box at the median redshift, and approximately 60 subboxes are replicated along the line of sight to reach this redshift. In this regime, the angular correlation function is largely underestimated at the scale of the catalogue (because it is larger than the scale of a box), and so the estimated cosmic variance is underestimated too.

Thirdly, it is interesting to consider the difference between the filled triangles and the open diamonds. At bright magnitudes, the two give the same variance. This is because the volume probed by the mock catalogues is much smaller than the volume of a subbox, so variance is well estimated with both methods, and is in fact Poissonian. At intermediate magnitudes, the volume probed by the mock cone is smaller than the full simulated volume, yet larger than that of a subbox. Hence, the subbox variance saturates at higher values. At the faint end, the lightcone is larger than the full simulated volume, so the variance shown with filled triangles suffers from a similar negative bias to that shown by open triangles. The three regimes are spanned here and in practice show that robust estimates of cosmic variance require a simulated volume much larger than the volume probed by the mock catalogue.

**Table 2.** Geometry of mock catalogues used for comparison with data from the APM and the K20 (see the text), and to make  $K_s$ -band counts. All mock catalogues have a square surface, of area given in the second column. The third column states how galaxies are selected in these catalogues (in the ‘Counts’ case, no photometric selection is applied, but the catalogues are truncated at a comoving distance of  $5500 h^{-1}$  Mpc from the observer). The fourth column gives the mean and span of the redshift distribution. The fifth column gives the mean and range of outputs used (output 70 is  $z = 0$ ). The sixth column gives the angular size of the simulated volume (of side  $L_b = 100 h^{-1}$  Mpc) at the mean redshift of the sample. The final column gives the number of radial replications needed to reach the mean redshift using the full simulated volume.

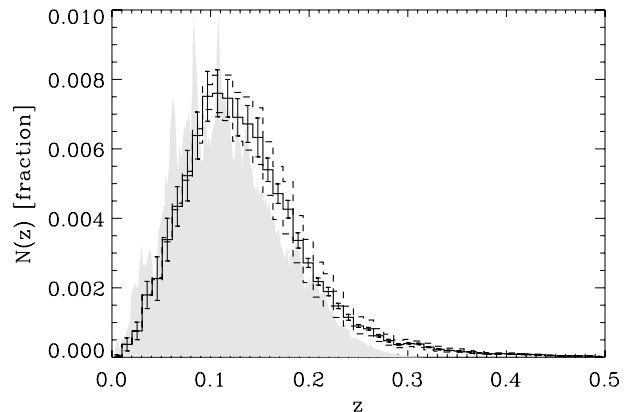
Sample	Area	Selection criteria	$\langle z \rangle$ ( $z_{\min}; z_{\max}$ )	$\langle i \rangle$ ( $i_{\min}; i_{\max}$ )	$\theta_{100}$ (deg)	$N_{100}$
APM	100 deg <sup>2</sup>	$17 < B_J < 20$	0.18 (0.04; 0.73)	65.3 (54; 69)	~10	~5.5
K20	1 deg <sup>2</sup>	$K_s < 20$	0.53 (0.02; 1.74)	58.0 (41; 69)	~4.5	~12
Counts	1 deg <sup>2</sup>	$\chi < 5500 h^{-1}$ Mpc	–	–	–	–



**Figure 9.** Effect of finite volume on the redshift distribution of a flux-limited sample. The solid (respectively, dashed) curve shows the mean redshift distribution estimated from 20 mock catalogues made from the full simulated volume (respectively, 125 mock catalogues each made from one different subbox). The attached error bars give the standard deviations (the small ones correspond to the solid curve). The upper panel gives the differential distributions while the lower panel gives the normalized cumulative distributions. The shaded area and dotted line show the locus of the K20 data in the top and bottom panels (see the text). The agreement with observations is very good.

Redshift distributions are affected by finite-volume effects in two ways. First, the variance and mean of the redshift distributions will change with box size. This effect is the same as that described above for the counts. Secondly, because the smaller the box is, the more replications are involved, repeated structures may imprint periodic features in  $N(z)$ . Thanks to the random tiling technique, this problem is avoided. In Fig. 9, we show the differential (upper panel) and cumulative (lower panel) redshift distributions of galaxies selected as in the K20 survey (Cimatti et al. 2002a). The shaded area (respectively, dotted curve) indicates the locus of the data in the upper panel (respectively, lower panel). In both panels, the curves (which are mostly overlaid) show the mean distributions measured from 20 mock catalogues made with the full simulated volume and from 125 mock catalogues each made with a different subbox. The error bars show the standard deviations about these means (the larger error bars correspond to the subbox catalogues). Fig. 9 shows that the redshift distribution does not change in shape when the size of the box varies. Although the details of  $N(z)$  will differ from one catalogue to the other, the statistical significance of redshift distributions is thus robust, and found to be in good agreement with  $K$ -band observations.

The agreement found with the K20 redshift distribution is an important success of the GALICS model, given the difficulty experienced by other models in achieving this task. The redshift distribution observed in the 2dFGRS (e.g. Colless et al. 2001) also seems to have been challenging for modellers to reproduce, and it is interesting to see how GALICS and MOMAF pass this test. The shaded area in Fig. 10 shows the redshift distribution of 2dF galaxies given by Colless et al. (2001). This distribution includes the whole survey, and thus corresponds to a nominal magnitude cut at  $b_J = 19.45$ . Because of various sources of incompleteness, however, the effective magnitude cut is more likely to lie around  $b_J = 19.3$  (see fig. 14 from Colless et al. 2001). The solid histogram in Fig. 10 shows the average redshift distribution measured in 20 mock surveys of  $10 \times 75 \text{ deg}^2$ , limited in apparent magnitude at  $b_J = 19.3$ . The associ-



**Figure 10.** Comparison of GALICS to the 2dFGRS redshift distribution. The shaded area shows data from Colless et al. (2001). The solid histogram (respectively, lower and upper dashed histograms) shows the GALICS redshift distribution for galaxies brighter than  $b_J = 19.3$  (respectively, 19.2, 19.4) estimated from 20 mock surveys of  $10 \times 75 \text{ deg}^2$ . The error bars show the dispersion in the estimates. The same arbitrary normalization was applied to the three histograms.

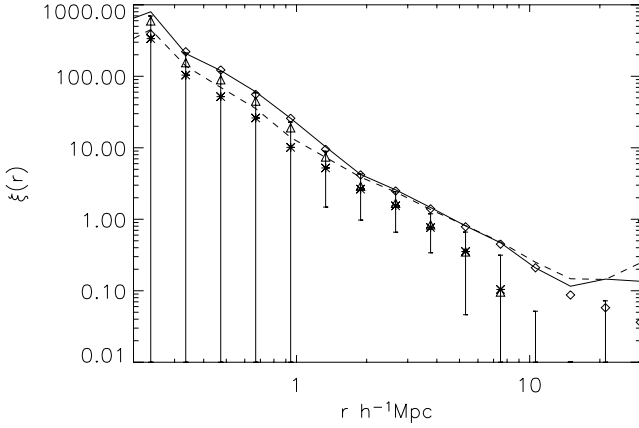
ated error bars show the dispersion around this mean. The dashed lines show the redshift distributions corresponding to an apparent magnitude cut at  $b_J = 19.2$  (lower line) and  $b_J = 19.4$  (upper line). The comparison suggests that the evolution of the  $b_J$ -band luminosity function predicted by GALICS is incorrect, giving too many bright galaxies at high redshifts. Indeed, an apparent magnitude cut at  $b_J \sim 19.1$  is necessary to bring our redshift distribution into better agreement with the 2dFGRS results. Let us note, however, that the scope of this comparison is limited in several ways. First, one should include a proper description of the complex selection function of the 2dFGRS for a more meaningful comparison. Although beyond the scope of this paper, this is readily feasible by applying the masks of the 2dFGRS to MOMAF mocks with a similar geometry. Secondly, the dispersion shown in Fig. 10 tells us that despite the huge amount of data gathered by the 2dFGRS, the cosmic variance is still quite large. Following the above discussion on finite-volume effects, and looking back at Fig. 3, one sees that this dispersion is bound to be an underestimate of the true cosmic variance because of the many replications involved in 2dFGRS-like mock surveys. This tells us that we need a bigger simulated volume to actually constrain the model: one needs to have a realistic cosmic variance at the size of a survey before one can hope to discriminate between different models. Finally, this example shows how useful the MOMAF software is for carrying out detailed comparisons of models with various data sets.

#### 4.2.2 Effects on estimates of two-point correlation functions

Finite-volume effects alter correlation functions in a complex way. Let us first discuss what happens to the spatial correlation function in a cubic box such as the simulated volume. The situation in mock catalogues is analogous but also includes projection effects. Following LS93 we relate the correlation function  $\hat{\xi}$  contained in the simulated volume  $V$  to the ‘real’  $\xi$  as

$$1 + \hat{\xi} = \frac{1 + \xi}{1 + \bar{\xi}_V}. \quad (16)$$

At small separations (compared with  $V^{1/3}$ ), where  $\bar{\xi} \ll \xi$ , the bias is negligible. At large scales,  $\bar{\xi} \sim \xi$  and  $\hat{\xi}$  falls down to 0. This bias



**Figure 11.** Spatial correlation function of galaxies brighter than  $B = -19$ . The diamonds show  $\xi_{100}$ , computed from the whole simulation, and the stars show the mean of estimates of  $\xi_{20}$  from 125 subboxes of side  $L_{\text{sb}} = 20 h^{-1}$  Mpc (stars). The error bars show the standard deviation about this mean. The dashed line shows  $\xi_{20}$  corrected for finite-volume effects (see the text), and agrees with the diamonds at large separations.

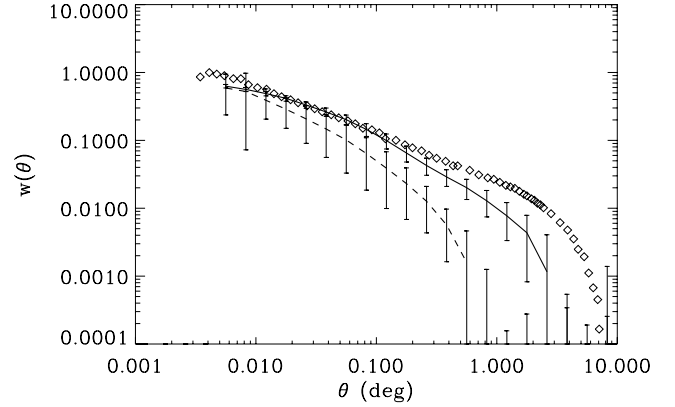
directly results from the fact that the variance cannot be estimated properly at the simulated volume scale, from only one simulated volume. Let us carry out numerical tests to better understand the finite-volume bias on the spatial correlation function. We again cut our standard simulation of side  $L_b = 100 h^{-1}$  Mpc into 125 cubic subboxes of side  $L_{\text{sb}} = 20 h^{-1}$  Mpc, and we measure the spatial correlation function in all of these 126 boxes, for galaxies brighter than  $B = -19$ . This magnitude cut leaves us with approximately 150 galaxies per subbox. In Fig. 11, we plot the SCF measured from the full simulation ( $\xi_{100}$ ) with diamonds, and the average of the 125 measures on subboxes ( $\langle \xi_{20} \rangle = \sum \xi_{20,i} / 125$ ) as stars. The error bars attached to the stars show the standard deviation from the 125 estimates of  $\xi_{20}$ . Comparison of  $\xi_{100}$  and  $\langle \xi_{20} \rangle$  shows that finite-volume effects translate into a negative bias at all scales, with a rather sharp cut-off at  $r \sim L_{\text{sb}}/5$ . The dashed line shows  $\langle \xi_{20} \rangle$  corrected from the integral constraint given in equation (16). The agreement of the dashed line with the diamonds is very good at large scales. At separations smaller than  $r \sim 1 h^{-1}$  Mpc, subbox-to-subbox fluctuations (both due to sparse sampling and to clustering) are responsible for the remnant discreteness bias. To understand this, let us consider the pair-weighted average of  $\xi_{20}$ :

$$\bar{\xi}_{20} = \frac{\sum_i n_i^2 \xi_{20,i}}{\sum_i n_i^2}, \quad (17)$$

where  $\xi_{20,i}$  is the correlation function measured (with the estimator from LS93) on the  $n_i$  galaxies of subbox  $i$ . This weighted average is shown with the triangles in Fig. 11. Note that in the case where edge effects are negligible (i.e. at small separations), one finds

$$\bar{\xi}_{20} \simeq \frac{\sum_i \text{DD}_i - 2\text{DR}_i + \text{RR}_i}{\sum_i \text{RR}_i} \simeq \frac{\text{DD} - 2\text{DR} + \text{RR}}{\text{RR}}, \quad (18)$$

where DD, DR and RR are the numbers of data–data, data–random and random–random pairs in a given separation bin for the whole simulated box. In other words, the pair-weighted average of the correlation functions of the subboxes is equivalent, at small scales, to using the LS93 estimator for the whole box, and is thus only affected by the integral constraint. Now, the main difference between this estimate and the estimate obtained from the full simulation (open diamonds) is that cross pairs between two subboxes are not regarded. In particular, as expected, this estimator converges to the



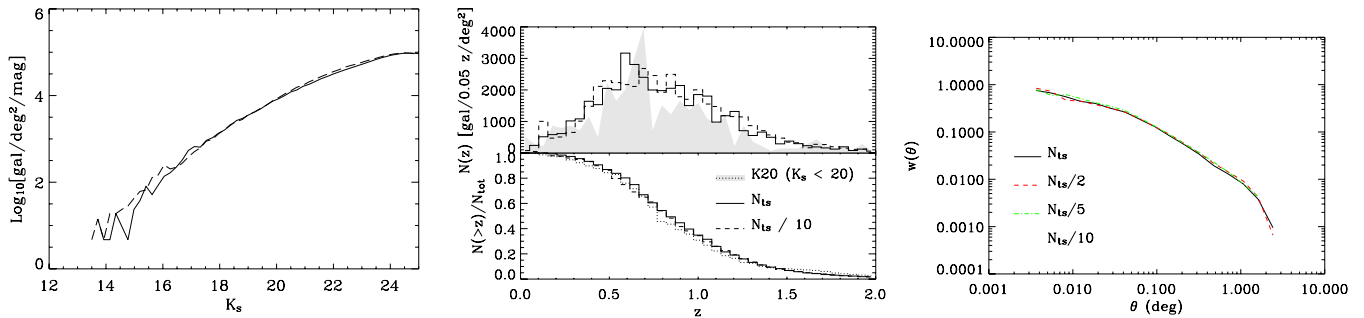
**Figure 12.** The solid (respectively, dashed) curve shows the mean ACF estimated from 20 mock catalogues made from the full simulated volume (respectively, 125 mock catalogues each made from one different subbox). The error bars give the standard deviations about these means. The cut-off in the subbox-catalogue ACF occurs at approximately one-fifth of the angular size of a subbox taken at the mean redshift of the galaxy sample. Open diamonds show the ACF measured from the APM (Maddox et al. 1996).

biased estimator ( $\xi_{20}$ ) at large scales. However, at small scales,  $\bar{\xi}_{20}$  partly captures subbox-to-subbox fluctuations through the variations of  $n_i^2$ , and thus remains above  $\langle \xi_{20} \rangle$ . Still,  $\bar{\xi}_{20}$  is a combination of estimates of the SCF on small boxes, which are contaminated at all scales by the integral constraint effect. Hence  $\bar{\xi}_{20}$  remains below the ‘exact’ result. When  $\bar{\xi}_{20}$  is corrected from the integral constraint as in equation (16) (solid line in Fig. 11), the result matches  $\xi_{100}$  nearly perfectly, as expected.

Fig. 12 shows how these finite-volume effects affect the angular correlation function. The solid line shows the mean of angular correlation functions measured from 20 mock automatic plate measurement (APM) catalogues (see Table 2) made from the full simulated volume, the attached error bars give the measured standard deviation. The dashed line and corresponding error bars show the mean and standard deviation of the ACF measured from the 125 mock APM catalogues made from the subboxes.

For the dashed line, the departure from a power law at large scales reported in the 3D case (see Fig. 11) occurs here at  $\theta \sim 0.4$ . This is a direct consequence of the finite volume of the subbox, and here 0.4 is approximately one-fifth of the angular size of a subbox at the median redshift of the survey (see Table 2). On top of this turn-around, there is an overall bias which increases slowly with separation starting at scales of approximately one-hundredth the size of a subbox. This is due to the projection of the bias described above for the SCF. Now, the open diamonds in Fig. 12 show the ACF measured from the APM by Maddox, Efstathiou & Sutherland (1996). The data are in very good agreement with our (full-box) model at scales shorter than  $\sim 0.1$ , which is approximately  $L_b/100$  at the median redshift. Longwards of this scale, the finite volume biases our ACF progressively. The comparison of our full-box  $w(\theta)$  to data from the APM is similar to the above comparison between subbox and full-box ACFs. We thus understand that the large-scale disagreement between APM and our model is not physical, but due to finite-volume effects: APM data are drawn from an even larger box, the Universe!

Finally, let us return to the issue of variance of counts in mock catalogues. Remembering that the variance of counts is basically given by the average of the angular correlation over the survey, we now clearly see how  $\sigma_3$  of the previous section was underestimated.



**Figure 13.** Left-hand-side panel: number counts from a mock catalogue using all available time-steps (solid curve) and one snapshot out of 10 (dashed curve). Middle panel: redshift distributions using all snapshots (solid histograms) or one out of 10 (dashed histogram). The data from the K20 survey are shown by the grey area in the upper panel and by the dotted line in the lower panel. Right-hand-side panel: angular correlation functions in catalogues containing all the snapshots (solid curve), one out of two (dashed curve), one out of five (dot-dashed curve) and one out of 10 (three-dot-dashed curve).

And we understand that this underestimation will occur unless we use a simulated volume more than 10 times larger than the aperture of the lightcone at the redshift of interest.

### 4.3 Other effects

#### 4.3.1 Time-step

The fact that we use a finite number of outputs, typically spaced in time by 100 Myr, could affect mock catalogues. We argued in Section 3.2.2 that this was not expected because even though individual galaxies can undergo a dramatic evolution during a time-step, the average properties (and their dispersion) of the overall population evolves at a much slower pace. Nevertheless, we check this hypothesis in this section by comparing statistics from mock catalogues made using different time-steps of the same simulation. Namely, we compare the counts, redshift distributions and ACFs obtained with our reference mock catalogues to those obtained with catalogues made using one snapshot out of 10.<sup>2</sup>

The resulting counts, redshift distributions and ACFs are shown in Fig. 13, and show no significant difference between the fine and coarse time-steps. This shows that the random tiling method is robust in that the resulting mock catalogues do not depend on the time-step used in the root simulation, provided the physics was properly integrated. The fact that the properties of mock catalogues do not change with the time-step shows that, at least for the selected galaxies, the  $K$ -correction and possibly a slow evolution determine the statistics. This justifies a posteriori the first-order correction made for the magnitudes in equation (4).

The necessity of using a fine time-step to make mock observations then mainly arises from the complex analysis that can be made from them (see Blaizot et al. 2004). In this perspective, one wants to retrieve the physical properties of individual galaxies and their hierarchical evolution for samples selected according to observational criteria. A short time-step naturally allows one to analyse evolution in more detail.

#### 4.3.2 Mass resolution

As mentioned in Section 2.3, the mass resolution of the DM simulation affects galaxies in three ways: (i) incompleteness, (ii) redshift

limit and (iii) ‘immaturity’. These three limitations, inherent to the hybrid method implemented in GALICS, will have different effects on the statistics measured from mock catalogues.

(i) Incompleteness sets in when a fraction of the galaxies of a given sample are missed because they would lie in haloes below the mass resolution of the DM simulation. This effect obviously causes underestimates of the counts at faint magnitudes. A more subtle effect is that a sample of galaxies affected by incompleteness will have a halo mass distribution biased towards high masses. Because more massive haloes are more clustered, this will induce a positive bias on correlation functions. These effects cannot be corrected for except by using simulations with a better mass resolution. However, as shown in Blaizot et al. (2004), it does not prevent one from using mock catalogues for studying bright galaxies, even at high redshift.

(ii) The redshift limit is the redshift beyond which no halo can be detected in a DM simulation. All possible galaxies at higher redshifts are thus missed by GALICS, and are hence missing from our mock catalogues. This effect, combined with incompleteness is responsible for a faint-end decrease in the counts.

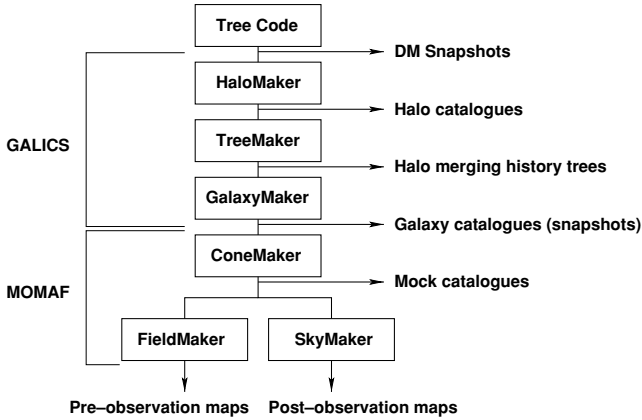
(iii) Immaturity describes the fact that young galaxies have unrealistic properties mainly because the cooling of gas in their host haloes was not slowed down by DM accretion in subresolution progenitors. These galaxies only become a significant part of the overall population at redshifts higher than  $\sim 2$  in our standard simulation, and they can be easily flagged and removed from a sample in our data base. In terms of apparent magnitudes, they only significantly affect  $K$ -band counts at  $K > 24$ .

The natural, and foreseen, solution to these limitations is to increase the resolution of the root DM simulations. We return to this perspective in the conclusions.

## 5 DATA BASE AND WEB INTERFACE

The current implementation of the GALICS hybrid model of hierarchical galaxy formation is a package that includes three main routines (see Fig. 14). First, HALOMAKER identifies haloes in each of the output snapshots. Secondly, TREEMAKER constructs the halo merging history trees from the list of haloes in all of the snapshots and computes the dark matter properties of each of the haloes. Thirdly, GALAXYMAKER deals with the fate of baryons within the merging history trees. It computes the properties of hot gas in haloes, and follows galaxy formation and evolution. The outputs are a list of properties (including absolute magnitudes in standard photometric bands) and rest-frame spectra for all galaxies in snapshots. The information

<sup>2</sup> Note that in any case, the properties of the galaxies were computed using all time-steps, which is necessary in order to properly describes the physics at stake in galaxy evolution (see Hatton et al. 2003).

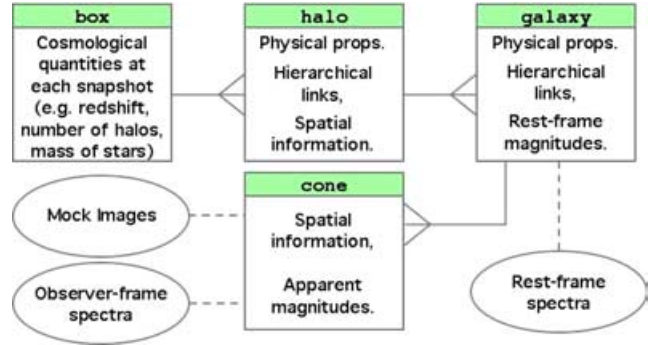


**Figure 14.** Organization chart of the GALICS/MOMAF project.

produced by a given GALICS post-processing of the simulation (defined by the choice of the astrophysical free parameters) constitutes what we hereafter call the archives of this post-processing (see GALICS I). Any change in the list of input parameters will correspond to a new issue of the post-processing and new output archives.

The MOMAF post-processing is also a package with three main routines, detailed in previous sections (see Fig. 14). First, CONEMAKER generates an observing cone by integrating along the light-cone through the various snapshots, and by managing the radial and transverse replications. The output is a list of galaxies with apparent properties (including apparent magnitudes in standard photometric bands) that are computed from the archives. Secondly, FIELDMAKER builds mock images from the observing cone, by projecting the cone galaxies on to the plane of the sky. Thirdly, any instrument simulator can be used to transform these pre-observation images into post-observation images, or the rest-frame spectra of the archives into post-observation observer-frame spectra. We have shown in the previous section how SKYKAKER can be used in this way, because FIELDMAKER is able to generate the relevant information in the proper format. Clearly, many different cones can be generated within a single GALICS archive by changing the observing point, the direction of the line of sight and/or the aperture. Many pre-observing images can be generated from a single cone by changing the filter response curves. And many post-observing images can be generated from a single pre-observing image by changing the instrument simulator. The information produced by a given MOMAF post-processing constitutes what we hereafter call the products of this post-processing.

At this stage, from any single simulation, we have generated a set of archives and products that includes tables of halo and galaxy properties, and FITS files of spectra and images. Two big issues obviously appear. First, the size of the data base makes it very cumbersome. As an example, for a standard GALICS post-processing of our  $\Lambda$ CDM simulation, the total numbers of haloes and galaxies generated in the 70 snapshots, respectively, amount to  $1.5 \times 10^6$  and  $1.8 \times 10^6$ . The output archives are approximately 4.5 GB for tables and 45 GB for spectra FITS files, not to speak of the products. Secondly, the specific information that is relevant for a given user is hidden within the bulk of non-relevant information. Let us imagine for instance that we want to obtain the  $B$ -band absolute magnitude and total cold gas mass of a random sample of 100 galaxies brighter than apparent magnitude  $I_{AB} = 20$ . Extracting this information will require reading tables with many columns (the properties) and many rows (the galaxies). It may be possible to anticipate the latter issue, and generate many specific tables for many different situations and



**Figure 15.** Data model of the GALICS/MOMAF data base. There are four layers of information corresponding to four tables. The box table contains information on mean quantities computed for each time-step, e.g. redshift, cosmic time and mean SFR. The halo table contains information on the DM haloes of each time-step, including their dynamical properties and merging history. The galaxy table contains information concerning the physical properties of galaxies at each time-step, including rest-frame magnitudes and spectra. The cone table contains information on the mock observing cone, mainly positions and apparent magnitudes. Companion FITS files include rest-frame spectra, pre- or post-observing images and observer-frame spectra.

potential users, but the same pieces of information will consequently be duplicated many times, which is not the proper way to proceed.

The solution to this conundrum is well known: it consists in storing the information into a relational data base (hereafter RDB). Here, we use the word data base (with its loose meaning) for all the information we want to make available and the word relational data base (with its strict meaning) for the technical way of putting part of this information into a specific structure.

We decided to use MySQL as the relational data base server. MySQL is a freely available, widely used and extremely fast data base server which is capable enough for our purposes. Tools to provide Web-based access to the MySQL server are also available. The tables generated by the GALICS and MOMAF post-processing are stored in MySQL tables. Our data base input and testing is performed using several short scripts in Perl that use the Perl/DBI module. The Web front-end uses PHP4 to pass SQL queries to the MySQL data base. Query outputs can either be displayed as an HTML table within the browser or downloaded to a local file. In this section we briefly describe the data base. A quick-start guide, sample queries and descriptions of the various fields in each table are available at the GALICS web-site (<http://galics.iap.fr>).

From a single dark matter simulation, each choice in the list of the input parameters corresponds to a GALICS post-processing with its specific archives and products, which in turn correspond to a single MySQL data base. The information is stored into a structure designed after the usual analysis in terms of entities, attributes and relationships, that is designed to minimize storage space and maximize query speed. Each MySQL data base is consequently organized in three MySQL tables for the archives, respectively, called the box, halo and galaxy tables, and numerous cone tables for the products. The data base scheme is illustrated in Fig. 15.

(i) The box table contains general information concerning mean quantities at each snapshot of the simulation, such as the cosmic time, the corresponding redshift, the total number of haloes and galaxies within the box, and integrated cosmic quantities such as the cosmic star formation rate, the cold gas content, the hot gas content, etc.

(ii) The halo table contains information on the haloes at each time-step. Each halo is identified by a unique ID in the simulation. This information deals with dark matter (e.g. the mass of the halo, the virial radius and the circular velocity) and the baryonic content of the haloes (e.g. the mass of hot gas and its metallicity). On top of this, we include spatial information, namely positions and velocities of the centres of mass of the haloes, and hierarchical information, i.e. merging history links. This information is, in principle, enough for one to run one's own semi-analytic model on our dark matter simulations, and thus freely test new recipes and compare results with GALICS.

(iii) The galaxy table contains the physical information we compute for galaxies: stellar masses, star formation rates, gas contents, rest-frame absolute magnitudes in a variety of filters, etc. Each galaxy is identified by a unique ID in the simulation.

(iv) The cone tables contain the positions of galaxies distributed in a mock catalogue, along with their apparent magnitudes in a variety of filters. Each galaxy is identified by a unique ID in the cone. However, because of transverse replication, different cone galaxy IDs can point to the same galaxy ID in the simulation. There are several cone tables, corresponding to different random seeds for the box shuffling process (which mimics, to some extent, cosmic variance) or to different field sizes.

Of course, the information included in the four tables is usable simultaneously in the queries, since the IDs of haloes and galaxies are shared by the tables of the data base and allow one to pass information from one table to another. Companion information on mock spectra and images is stored as FITS files. The rest-frame spectra are related to the galaxies in the galaxy table, whereas the images are related to a particular cone table. The observer-frame spectra are related both to galaxies and to the cone from which the galaxies are identified.

The GALICS web-site also contains a hierarchical query page which allows the user to retrieve hierarchical information for any galaxy in mock catalogues or snapshots. At the moment, this page contains three scripts. The first one allows one to view the full merging tree of a given galaxy, identified by its unique ID. The second one allows one to follow the evolution of selected properties of a galaxy along its merging history tree. Here, three options are available: (i) one can follow properties along the main branch which links a galaxy to its most massive progenitor at each merger; (ii) one can follow the most massive branch, which links the most massive progenitors at each time-step together; or (iii) one can perform a sum of additive properties on all progenitors at each time-step. Each of these options is a different way of retrieving partial information contained in the full merging history tree of a galaxy. The third script allows one to obtain the list of descendants or progenitors (at any redshift) of a sample of galaxies selected with any set of criteria. An example of use of this powerful script can be found in Blaizot et al. (2004). These scripts allow, for the first time, the systematic exploration of the evolution of galaxies within the framework of hierarchical galaxy formation.

## 6 CONCLUSIONS

In this paper, we presented the Mock Map Facility that takes the results of our GALICS hybrid model of hierarchical galaxy formation to make mock galaxy samples. Our method involves the construction of observing cones by integrating through the snapshots of the  $N$ -body simulation, and by using the properties of galaxies as they are computed by the GALICS post-processing. This technique builds

up on the simulation and is affected by the shortcomings of the latter (mass resolution, and the absence of rare objects due to the limited size of the box). It also incorporates shortcomings due to radial replication along the line of sight, and, for large solid angles, transverse replication. We introduced box reshuffling to minimize replication effects. The price of this technique is the loss of some signal for the correlation functions [both two-dimensional (2D) and 3D] on distances smaller than the size of the box. This loss is generally not larger than 10 per cent. Of course, there is no signal on distances larger than the size of the box, and finite-volume effects have been shown to introduce a significant (but well understood) bias on angular correlation functions.

For the purpose of analysing the limitations of our method, we compared predictions of GALICS/MOMAF to various observations. We showed that the model agrees well with  $K$ -band counts and redshift distributions. And we showed that within finite-volume effects, the model also agrees well with the APM angular correlation function. These results have been obtained with the same model as used in Blaizot et al. (2004) which showed good agreement with the properties of Lyman break galaxies at  $z \sim 3$ . This shows that our mock catalogues can readily be used for a variety of scientific investigations.

From the mock catalogues of the observing cones, we show how to make 'realistic' mock images. Since our GALICS post-processing involves multi-wavelength information from the ultraviolet (UV) to the submillimetre range, our mock images are produced through a wide range of standard filters. These field images can be observed through any instrument simulator. The technique is able to produce input lists for SKYKAKER. Instrument simulators adapted to observations at infrared and submillimetre wavelengths can also be used.

The data base produced by the GALICS and MOMAF post-processing is quite large, and has to be stored in such a way that easy access to relevant information is provided. We put the results into a relational data base structure to which SQL queries can be passed through a simple Web interface. This structure has a number of well-known advantages: it optimizes storage space, it makes access to the relevant information very easy, it is able to deal with simultaneous queries and updates, etc. The results of GALICS (physical properties, rest-frame magnitudes) and MOMAF (observable properties, apparent magnitudes) are stored in this data base, and linked together through the standard system of a relational data base model. FITS files of mock images and spectra are also available linked from the data base.

The content of the data base can be used for several purposes. For instance:

(i) comparison of mock predictions with observations through the production of a mock survey that can be processed with the same data processing pipeline as the actual survey;

(ii) elaboration of observing strategies for forthcoming satellite missions and ground-based instruments;

(iii) benchmark for data processing pipelines; a data base populated with GALICS sources is a valuable 'test set' on which to base and test the various techniques and algorithms for data reduction and analysis, for the next generation of astronomical instrumentation; the data base includes the positions and magnitudes of the galaxies that are put into the mock images, and can be used as a 'truth table' that has to be recovered by the data processing software;

(iv) creation of customized galaxy samples for comparison with other models or observational data.

We are considering improvements to this prototype data base. They can develop along three axes. (i) In the mid-term future, the



foreseen level of computer performance makes replications unavoidable if a sufficient level of mass resolution for galaxy studies has to be attained. However, the improvement of the simulations will result in larger boxes that will decrease the number of radial and transverse replications, and be able to include rarer objects. (ii) The improvement of the physics within the simulations will also make better mass resolution possible, and will (hopefully) produce better results. There is also no doubt that the semi-analytic recipes have to be improved. The same cone building technique will also be used for converting the outputs of  $N$ -body simulations + hydrodynamics into mock observations. (iii) The data base prototype that has been presented here will be enhanced to make it compatible with the data and metadata standards that are now being developed as part of the theoretical virtual observatory. The present MOMAF will form a valuable test-bed for testing the integration of theoretical data from simulations into the theoretical virtual observatory, which forms a part of the global Astronomical Virtual Observatory effort.

## ACKNOWLEDGMENTS

The authors thank the anonymous referee for insightful comments that helped to improve this paper. YW was supported by projects 1610-1 and 1910-1 of the Indo-French Centre For Advanced Scientific Research (CEFIPRA). The  $N$ -body simulation used in this work was run on the Cray T3E at the IDRIS supercomputing facility.

## REFERENCES

- Benson A. J., Cole S., Frenk C. S., Baugh C. M., Lacey C. G., 2000, *MNRAS*, 311, 793
- Berlind A. A., Weinberg D. H., 2002, *ApJ*, 575, 587
- Bernardeau F., Colombi S., Gaztanaga E., Scoccimarro R., 2002, *Phys. Rep.*, 367, 1
- Bertin E., Arnouts S., 1996, *A&AS*, 117, 393
- Blaizot J., Guiderdoni B., Devriendt J. E. G., Bouchet F. R., Hatton S. J., Stoehr F., 2004, *MNRAS*, 352, 571 (GALICS III)
- Bouwens R., Broadhurst T., Silk J., 1998, *ApJ*, 506, 557
- Cimatti A. et al., 2002a, *A&A*, 391, L1
- Cimatti A. et al., 2002b, *A&A*, 392, 395
- Coil A. L., Davis M., Szapudi I., 2001, *PASP*, 113, 1312
- Cole S., Hatton S., Weinberg D. H., Frenk C. S., 1998, *MNRAS*, 300, 945
- Colless M. et al., 2001, *MNRAS*, 328, 1039
- Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, *ApJ*, 292, 371
- Devriendt J. E. G., Guiderdoni B., Sadat R., 1999, *A&A*, 350, 381
- Diaferio A., Kauffmann G., Colberg J. M., White S. D. M., 1999, *MNRAS*, 307, 537
- Eke V. R., Cole S., Frenk C. S., 1996, *MNRAS*, 282, 263
- Erben T., Van Waerbeke L., Bertin E., Mellier Y., Schneider P., 2001, *A&A*, 366, 717
- Evrard A. E. et al., 2002, *ApJ*, 573, 7
- Górski K. M., Banday A. J., Hivon E., Wandelt B. D., 2002, in *ASP Conf. Ser.* 281: *Astronomical Data Analysis Software and Systems XI HEALPix – a Framework for High Resolution, Fast Analysis on the Sphere*. Astron. Soc. Pac., San Francisco, p. 107
- Hatton S., Devriendt J. E. G., Ninin S., Bouchet F. R., Guiderdoni B., Vibert D., 2003, *MNRAS*, 343, 75 (GALICS I)
- Helly J. C., Cole S., Frenk C. S., Baugh C. M., Benson A., Lacey C., 2003, *MNRAS*, 338, 903
- Hernquist L., 1990, *ApJ*, 356, 359
- Kauffmann G., Nusser A., Steinmetz M., 1997, *MNRAS*, 286, 795
- Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1999, *MNRAS*, 303, 188
- Kennicutt R. C., 1983, *ApJ*, 272, 54
- Landy S. D., Szalay A. S., 1993, *ApJ*, 412, 64 (LS93)
- Madau P., 1995, *ApJ*, 441, 18
- Maddox S. J., Efstathiou G., Sutherland W. J., 1996, *MNRAS*, 283, 1227
- Ninin S., 1999, PhD thesis, Université Paris 11
- Peacock J. A., Smith R. E., 2000, *MNRAS*, 318, 1144
- Scoccimarro R., Sheth R. K., Hui L., Jain B., 2001, *ApJ*, 546, 20
- Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001, *MNRAS*, 320, 289
- Steidel C. C., Hamilton D., 1993, *AJ*, 105, 2017
- Yang X., Mo H. J., Jing Y. P., van den Bosch F. C., Chu Y., 2004, *MNRAS*, 350, 1153

## APPENDIX: USING THE DATA BASE

In this appendix, we briefly illustrate how the data base can be used to interpret observational data in the paradigm of hierarchical galaxy formation. We give four examples which exemplify the kind of information that can be exploited:

- (i) synthesis of a volume-limited sample of galaxies, for instance at  $z \simeq 0$ ;
- (ii) synthesis of a magnitude-limited sample of galaxies, and related multi-wavelength information;
- (iii) search for 2D and 3D spatial information (e.g. for redshift distribution, clustering), and correlation of properties with it;
- (iv) search for merging history trees within hierarchical galaxy formation (e.g. in what type of galaxy does the material of a given high-redshift galaxy end up at  $z \simeq 0$ ? How many progenitors does a galaxy at  $z \simeq 0$  have?).

For each example, we give a typical SQL query that returns the requested subsample by querying the data base. We refer the reader to the Web page (<http://galics.iap.fr/>) for additional examples, and a simple introduction to SQL syntax.

### A1 Volume-limited samples

It is possible to query the data base to list a series of physical properties for a subsample of galaxies with sophisticated selection criteria. As an example, select 100 galaxies at random in the  $z = 0$  snapshot (that corresponds to a time-step of 70), with the requirement that their Johnson  $B$ -band absolute magnitude is brighter than  $-20$ , and their dispersion velocity is larger than  $200 \text{ km s}^{-1}$ . We are also interested in obtaining their absolute  $K$ -band magnitude,  $B - K$  colour, morphological types and total stellar mass.

```
> SELECT gal_id, type_B2D_lum, tot_JOHNSON_B,
> tot_speed, tot_JOHNSON_K,
> tot_JOHNSON_B-tot_JOHNSON_K, tot_mstar
> FROM galaxy
> WHERE time-step = 70
> AND tot_JOHNSON_B < -20
> AND tot_speed > 200
> ORDER BY RAND()
> LIMIT 100
```

The final two commands place the list of galaxies recovered by the query in random order, and then limit the output to the first 100 rows. This example takes less than 1 s to run.

### A2 Magnitude-limited samples

Another type of query is to select galaxies according to their apparent magnitudes in order to mimic an observational sample or to predict what a forthcoming survey will yield. The selected sample can then be studied in the explicit cosmological context of GALICS,

and the physical properties of the selected galaxies can be retrieved easily to gain insight into the nature of the ‘observed’ objects. An example of using mock catalogues to interpret observational data is given in Blaizot et al. (2004). A crucial issue in observational galaxy formation studies is the identification of counterparts at any wavelength of galaxies observed through any given filter. It is often quite a challenge, for example, to identify the optical counterparts of far infrared sources, observed with low angular resolution. The GALICS data base provides a powerful tool to address these questions as it predicts emission properties of galaxies from UV to submillimetre bands and gives the opportunity to build corresponding mock maps.

An example SQL query to retrieve a magnitude-limited sample in a  $1 \text{ deg}^2$  cone is given below, for galaxies brighter than  $I_{AB} = 22.5$ . We are interested in their apparent  $B - K$  colour (in the  $AB$  system), their total stellar mass and the virial mass of their host halo. Such a query requires information that is present not only in the cone table, but also in the galaxy table and the halo table. It requires what is called a join in SQL syntax:

```
> SELECT cone_001.cone_id, cone_001.app_redshift,
> cone_001.JOHNSON_BAB, cone_001.JOHNSON_KAB,
> cone_001.JOHNSON_BAB-cone_001.JOHNSON_KAB,
> galaxy.tot_mstar, halo.m_vir
> FROM galaxy, cone_001, halo
> WHERE cone_001.JOHNSON_IAB < 22.5
> AND cone_001.gal_id = galaxy.gal_id
> AND halo.halo_id = galaxy.halo_id
```

This example query runs in  $\sim 30$  s, and returns information (seven columns) for approximately 31 000 galaxies.

The same type of selection can be used to work the other way around: one can select galaxies according to their physical properties or their dark matter halo properties, and extract their spatial distribution and apparent magnitudes from the cone.

### A3 Spatial information

Spatial information can be retrieved the same way as above, for galaxies in mock catalogues. Consider the following query:

```
> SELECT right_ascension, declination, app_redshift
> FROM cone_001
> WHERE JOHNSON_IAB < 22.5
```

This produces a table with the angular coordinates and apparent redshifts of all the mock galaxies brighter than 22.5 in the  $I_{AB}$  band within a  $1 \text{ deg}^2$  field. There are again approximately 31 000 such galaxies, and the query here runs in  $\sim 5$  s.

### A4 Hierarchical evolution

The GALICS project gives one, for the first time, the opportunity to interpret observational data within the paradigm of hierarchical galaxy formation in a systematic way. One of the most important features of this theoretical framework is the notion of a galaxy merging history tree. Going up or down this tree allows one to investigate the properties of the progenitors or descendants of any given galaxy at any redshift, and the mass build-up of that galaxy. In the galaxy table, each galaxy has a pointer towards its unique descendant at the next time-step. This minimal information is sufficient to reconstruct any merging history tree, whether forward (the list of descendants, that is a single branch as time flows) or backwards (the list of progenitors, that may be a full tree with many branches as we look back). The number of merging events for the progenitors of the galaxy where the ID is (the character string) `xxyyyyyzzz` is easily obtained through the following query, and the ID of its descendant at the next time-step (for all time-steps except the final one):

```
> SELECT nb_merge, daughter_num
> FROM galaxy
> WHERE gal_id = 'xxyyyyyzzz'
```

It is necessary to run the above query recursively to build up the full merging history trees. On the GALICS website we provide PHP scripts that generate such recursive queries and pass them to the data base server. Once a galaxy ID is supplied by the user through the Web interface, the IDs of its progenitors and descendants, and their properties, are recovered through the ‘recursive query’ page. An interesting option allows the user to obtain the sum of the (additive) properties of all the progenitors. In such a way, the evolution of the total star formation rate or the total stellar mass in all the progenitors can be easily followed. Examples of such recursive queries can be found in GALICS II and GALICS III.

This paper has been typeset from a  $\text{T}_{\text{E}}\text{X}/\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  file prepared by the author.