Gravitational quenching by clumpy accretion in cool-core clusters: convective dynamical response to overheating

Yuval Birnboim\textsuperscript{1,2}\textsuperscript{*} and Avishai Dekel\textsuperscript{2}\textsuperscript{*}

\textsuperscript{1}Harvard Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
\textsuperscript{2}Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel

Accepted 2011 April 8. Received 2011 March 29; in original form 2010 August 4

ABSTRACT

Many galaxy clusters pose a ‘cooling-flow problem’, where the observed X-ray emission from their cores is not accompanied by enough cold gas or star formation. A continuous energy source is required to balance the cooling rate over the whole core volume. We address the feasibility of a gravitational heating mechanism, utilizing the gravitational energy released by the gas that streams into the potential well of the cluster dark matter halo. We focus here on a specific form of gravitational heating in which the energy is transferred to the medium through the drag exerted on inflowing gas clumps. Using spherically symmetric hydro simulations with a subgrid representation of these clumps, we confirm our earlier estimates that in haloes \(\geq 10^{13} \text{M}_\odot\) the gravitational heating is more efficient than the cooling everywhere. The worry was that this could overheat the core and generate an instability that might push it away from equilibrium. However, we find that the overheating does not change the global halo properties, and that convection can stabilize the cluster by carrying energy away from the overheated core. In a typical rich cluster of \(10^{14} - 15 \text{M}_\odot\), with \(~5\) per cent of the accreted baryons in a clump of \(~10^8 \text{M}_\odot\), we derive upper and lower limits for the temperature and entropy profiles and show that they are consistent with those observed in cool-core clusters. We predict the density and mass of cold gas and the level of turbulence driven by the clump accretion. We conclude that gravitational heating is a feasible mechanism for preventing cooling flows in clusters.

Key words: hydrodynamics – galaxies: clusters: general – galaxies: formation – galaxies: haloes – X-rays: galaxies: clusters.

1 INTRODUCTION

Galaxy clusters can be divided into two distinct populations according to the X-ray luminosity of their central cores (Sanderson, Ponman & O’Sullivan 2006). Cool-core (CC) clusters are centrally concentrated, highly luminous in X-ray and have central cooling times of 0.1–1 Gyr. Non-cool-core (NCC) clusters have lower densities and luminosities near the centre, and their central cooling times are typically a few Gigayears (Donahue et al. 2006). The population of CCs has little internal variability, and they all exhibit a typical temperature profile with a decline by a factor of 2–3 in the innermost few 10 kpc (Leccardi & Molendi 2008). Based on the short cooling time in CCs, one expects to observe gas in intermittent temperatures, a high star formation rate in the brightest central galaxy (BCG; \(\gtrsim 100 \text{M}_\odot \text{yr}^{-1}\)) and a large stellar mass in the BCG (\(10^{12} - 13 \text{M}_\odot\)), none of which is observed. These are three manifestations of the cooling-flow problem in clusters (Fabian 1994). Since the cooling time is inferred directly from observations based on the luminosity and temperature, the discrepancy between the expected cooling rates and the gas that actually cools indicates that some heating mechanism is keeping the gas hot in a stable configuration.

Several mechanisms have been proposed as potential solutions to the cooling-flow problem. Most popular is active galactic nucleus (AGN) feedback, where energy or momentum is provided by an active galactic nucleus either in an intense quasar mode (Ciotti & Ostriker 2007), in a slower radio mode (see Best 2007; Cattaneo et al. 2009, for a review) or via cosmic rays (Guo & Oh 2008). The AGN clearly release sufficient power to balance the cooling in the cores, and the observed big radio and X-ray bubbles in some cluster cores (Birzan et al. 2004) is likely evidence for AGN feedback. However, the coupling of the AGN energy to the gas in the whole core volume is not easily understood, and the requirement of continuous heating is not a trivial constraint (De Young, O’Neill & Jones 2008). Another possibility is that the cooling instability might be locally suppressed with the addition of non-thermal pressure sources such as cosmic rays or turbulent magnetic fields (Sharma, Parrish & Quataert 2010).

*E-mail: ybirnboim@cfa.harvard.edu (YB); dekel@phys.huji.ac.il (AD)
Here we utilize the fact that the gravitational power released as fresh baryons stream into the potential well created by a massive halo is more than enough to balance the cooling rate at the centre (Fabian 2003; Dekel & Birnboim 2008, hereafter DB08; Wang & Abel 2008). Among the mechanisms through which this energy is transferred to the inner halo one could consider dynamical friction (El-Zant, Kim & Kamionkowski 2004; Faltenbacher & Mathews 2007; Naab et al. 2007; Khochar & Ostriker 2008; Johansson, Naab & Ostriker 2009), thermal conduction (Kim & Narayan 2003) or turbulence (Sato & Nagataki 2004). In particular, the turbulence induced by accreting substructures may disturb the magnetic field in a way that could make the conduction more efficient (Parrish, Quataert & Sharma 2010; Ruszkowski & Oh 2010).

The gravitational power that is released during the streaming of baryons into clusters of $M \geq 10^{13} M_\odot$ is indeed sufficient for balancing the expected radiative losses (DB08). The challenge is to deposit this energy at the inner cluster core of $\sim 30–100$ kpc, where most of the cooling takes place. The energy should be evenly distributed over the whole core volume and continuously over several Gigayears. It should be performed in a way that is consistent with the observed entropy profile (Donahue et al. 2006) and metallicity (Rebusco et al. 2006) profile. Here we investigate a model in which some of the accreted gas is in dense and cold ($\sim 10^4 K$) gas clumps. These clumps do not stop at the virial shock but rather penetrate to the inner parts of the halo. The continuous accretion of many clumps through the halo can distribute the energy smoothly in a large volume, as required, unlike AGN feedback that is episodic and originates from a very small region near the black hole. The main physical mechanism that couples the clumps and the halo gas is hydrodynamic drag, subsonic or supersonic, which decelerates the clumps and causes the deposition of their kinetic and potential energy in the ambient medium. Hydrodynamic drag is more efficient when the clumps are smaller and moving faster, as opposed to dynamical friction that becomes more efficient for more massive clumps and for transonic velocities (Ostriker 1999). We pointed out in DB08 that the gravitational heating is likely to be associated with the streams that build the cluster along the filaments of the cosmic web (Dekel & Birnboim 2006; Dekel et al. 2009). These streams could transfer their energy to the cluster core via the generation of turbulence and other mechanisms that are not necessarily associated with cold clumps (Zinger et al., in preparation). Still, the heating through cold clumps is a concrete example that allows a simple study of the dynamical response with general implications that are not limited to this particular coupling mechanism.

In DB08 we showed that accreted clumps could balance the cooling in haloes more massive than $6 \times 10^{12} M_\odot$, provided that the clumps contain a non-negligible fraction of the infalling gas ($\geq 10$ per cent for a $10^{13} M_\odot$ halo), and that the clump masses are in the range $10^{6}–10^{8} M_\odot$. The clumps heat the intracluster medium (ICM) via drag until they disintegrate by hydrodynamical instabilities (Maller & Bullock 2004; Murray & Lin 2004). The permitted mass range for the clumps took into account additional criteria for clump survivability, such as Bonnor–Ebert instability, conduction and evaporation. When the clumps are sufficiently massive, they survive long enough to penetrate through the outer halo and reach the centre, while by being not too massive, their interaction with the ICM is sufficiently strong for most of the clump energy to be deposited in the core within a Hubble time. The basic features studied in DB08 assuming a static system will be implemented below in a dynamical evolving configuration.

The present study is motivated by a potential overheating problem. The heating rate by drag (and by other mechanisms such as dynamical friction, cosmic rays from AGN) is proportional to the density of the hot ambient gas, $\dot{e}_{\text{heat}} \propto \rho_{\text{hot}}$. On the other hand, the bremsstrahlung cooling rate per unit volume scales like $\dot{e}_{\text{cool}} \propto \rho_{\text{hot}}^{2.5}$ (assuming isobaric gas). This generates an instability through a positive feedback loop (Field 1965; Conroy & Ostriker 2008). A small negative density perturbation, e.g. produced by gas expansion due to overheating, would make the ratio of heating to cooling rate increase as $\dot{e}_{\text{cool}}/\dot{e}_{\text{heat}} \propto \rho_{\text{hot}}^{-3}$, leading to more overheating, enhanced pressure and a runaway expansion. An analysis of the consequences of this instability, and an investigation of possible mechanisms that could keep the cluster in equilibrium, necessitate a dynamical treatment.

This unstable overheating, as reproduced in sphericosymmetric simulations below, results in expanding shells that heat to temperatures as high as $10^9 K$, clearly at odds with observations. In the real world, this overheating must be regulated by processes that smooth temperature or entropy gradients by heat transfer through conduction or convection. Convection is suppressed in the presence of magnetic fields (Fabian 1994, and reference within), though it might be boosted almost to its maximum possible value (Spitzer 1962) by turbulence (Narayan & Medvedev 2001; Balbus & Reynolds 2008; Parrish et al. 2010; Ruszkowski & Oh 2010), which might be a natural product of clump heating. Convection is a promising mechanism for smoothing the local instabilities. In the simple case of an ideal gas with a uniform chemical composition in a spherical potential well, convection occurs in regions where the entropy is declining with radius. However, the strength of this convection is uncertain, as it depends on the gradients in gas properties and on the magnetic fields. Weak magnetic fields make the gas more susceptible to convection, with the entropy-gradient criterion replaced by the temperature gradient (Balbus 2000, 2001; Quataert 2008), but for certain types of perturbations, the convection strength might be drastically suppressed by effects related to magnetic tension (Parrish, Stone & Lemaster 2008; Parrish, Quataert & Sharma 2009). Regardless of the actual energy transport mechanism, one expects nature not to permit shells of $\sim 10^9 K$ in close contact with shells of $\sim 10^4 K$, and to act to smooth such a discontinuity. Motivated by the turbulence that is expected to be generated by the clumps, we focus below on convection as the mechanism that smooths steep gradients.

In this paper, we mimic this smoothing process by a 1D mixing length convection model (Spiegel 1963), with a the mixing length coefficient $\ell$, the single free parameter. We find that the results are almost independent of the value of this parameter as long as noticeable convection occurs. This allows us to further simplify the model by assuming that the convection is maximal, namely the energy transfer rate is limited by the requirement that hot bubbles accelerate until they become supersonic, at which point they disintegrate. This leaves us with no free parameters in our convection model.

There are some additional benefits from a dynamic treatment of the clump heating process. The analysis of DB08 considered simple Monte Carlo clump trajectories within an otherwise static halo in hydrostatic equilibrium. This assumption of a static halo could be valid for 1 Gyr but the cluster may evolve considerably over a Hubble time, due, for example, to the gradual increase of virial temperature and the growth of the BCG. Furthermore, if clump heating is taking place, cold clumps continually get destroyed near the halo centre, dumping cold gas near the BCG. This dilution of the hot gas with cold gas is not a problem for the heating–cooling
balance because the clumps bring in several times the energy needed for heating themselves to the cluster virial temperature, but a proper account of this clump deposition requires a dynamical analysis of an evolving cluster.

In Section 2 we describe the implementation of the clump model and of the convection model in the 1D hydrodynamic code. In Section 3 we show results of hydrodynamic cluster simulations with convection and clump heating that match the observed temperature and entropy profiles of clusters and the cooling rates in clusters without a need for any additional feedback. In Section 4 we address possible direct and indirect observations of the cold clumps. In Section 5 we summarize and conclude.

2 METHODS

2.1 Implementing clumps in 1D hydrodynamic simulations

According to our estimates in DB08, heating by clumps requires clump masses in the range $10^4$–$10^6 M_{\odot}$. In order to properly resolve drag forces and clump disintegration via hydrodynamical instabilities, these clumps should be resolved by at least 1000 cells or smoothed particle hydrodynamics (SPH; i.e. 10 cells across each dimension). The implied required dynamical range in a cluster of $10^{14}$–$10^{16} M_{\odot}$ is impractical, so simulation of such clumps requires a subgrid model. We develop such a model below, and describe its implementation in a 1D spherical hydrodynamical code.

The clumps are made of cold and partly ionized gas at $\sim 10^4$ K in pressure equilibrium with their surrounding hot halo. For a rich cluster of galaxies, with a virial temperature $\sim 10^8$ K, the overdensity within the clumps is about $10^3$. The clumps couple to the hot gas by a drag force

$$f_{\text{drag}} = \frac{1}{2} C_d A \rho_{\text{hot}} v_{\text{rel}}^2,$$

acting opposite to the relative direction of motion. Here $C_d$ is the drag coefficient ($\sim 1$ for a spherical gas cloud), $A$ is the cross-section surface area of the clump ($\pi R_{\text{cl}}^2$), $\rho_{\text{hot}}$ is the density of the hot component and $v_{\text{rel}}$ is the relative velocity between the clump and the halo gas. Equation (1) holds for subsonic and supersonic motions, though the value of $C_d$ may vary, especially in the trans-sonic regime where it could be a few. The deceleration ($f_{\text{drag}}/m_{\text{cl}}$) is proportional to the ratio of clump surface area to volume, $\propto m_{\text{cl}}^{1/3}$. This dictates lower and upper limits to the relevant clump masses. Clumps that are too small cannot penetrate through the outer halo into the core, and clumps that are too large cannot deposit a significant fraction of their energy in the inner halo on a time-scale shorter than the Hubble time.

Single clump simulations (Murray & Lin 2004) indicate that most of the energy dissipated in this process goes into the hot ambient gas. The survivability of these clumps is an open question as they could be destroyed by many different effects. This has been discussed in DB08 and in Maller & Bullock (2004). In a nutshell, if the clump is more massive than $\sim 10^5 M_{\odot}$ it would exceed the Bonnor–Ebert critical mass (the equivalent of the Jeans mass for pressure-confined spheres) and it would collapse under its self-gravity and turn into stars. If the clump is less massive than $\sim 10^4 M_{\odot}$, evaporation and conduction are expected to disintegrate the clump. Hydrodynamical instabilities, particularly Kelvin–Helmholtz instability, tend to break the clump, typically after it has expelled the equivalent of its own mass in ambient gas (Murray & Lin 2004). The clump masses then cascade down toward the lower limit for clump mass.

2.1.1 The HYDRA code

Subgrid recipes of the effects mentioned above were incorporated into the 1D spherical code HYDRA (Birnboim & Dekel 2003). The code is finite-difference Lagrangian with von-Neumann first- and second-order artificial viscosity. Dark matter is described as zero-width shells that propagate through the gaseous shells, and interact with them gravitationally. The coupled density fields of gas and dark matter shells are propagated using a fourth-order Runge–Kutta method. Time-steps are defined by the minimum of the Courant conditions and the allowed deviation of the fourth-order scheme from fully explicit (first order) time-step. When this difference exceeds some preset epsilon, the previous values are restored, and a new step with decreased time-step is performed. A comprehensive description of the hydrodynamic and dynamic equations for the gas and the dark matter shells, the numerical scheme used and convergence tests and a comparison to analytic test problems can be found in Section 3 and the appendixes of Birnboim & Dekel (2003).

The baryons and the dark matter shells are assigned a preset angular momentum that is added as an impulse when the shell is at its turn-around. This prevents a numerical and physical singularity at the centre. The angular momentum of the dark matter is set so that the rotational kinetic energy is 18 per cent of the radial kinetic energy at the virial radius. The results are insensitive to this choice. The baryonic angular momentum is set to produce a spherical, angular momentum supported ‘disc’ of radius $\sim 10 kpc$ for a cluster halo of $10^{15} M_{\odot}$, to mimic a BCG. Since centrifugal forces scale like $r^{-3}$, the angular momentum of the baryons is negligible at a distance comparable to a few disc radii above the disc. We note that a spherical code is not the right tool for studying disc formation, and this set-up is essentially an inner boundary condition for the halo simulation. In addition, the code imposes a central smoothing length on the gravitational acceleration, $a_g = GM(r + s)^2$, typically with $s \sim 50$ pc. Radiative cooling is calculated by a metallicity-dependent cooling function (Sutherland & Dopita 1993) with a constant preset metallicity.

The initial conditions, in terms of shell masses, radii and peculiar velocities, were set at $z = 100$ such that the future accretion rate on to the growing halo will follow a desired accretion rate (Dekel 1981; Birnboim & Dekel 2003, appendix C). Specifically, the initial perturbation used here yields an accretion history that traces that of an average main progenitor according to the extended Press and Schechter (EPS) approximation (Press & Schechter 1974; Lacey & Cole 1993; Neistein, van den Bosch & Dekel 2006). The procedure is described in detail in Birnboim et al. (2007). The code has been compared successfully to analytic predictions of Bertschinger (1985) and to a Von-Neumann–Sedov–Taylor problem. The Courant conditions and epsilons are set so that the global energy conservation over a Hubble time is always better than 1 per cent and the spatial convergence was tested for each set of simulations. Time-steps are consequently $\sim 10^{-5}$ Gyr throughout most of the simulation.

2.1.2 Drag forces in 1D

The subgrid model for clumps is similar in its approach to ‘sticky particle’ techniques in the sense that it calculates subgrid interactions on otherwise N-body particles. Here we define ‘clump-shells’, which, like dark matter shells, are able to penetrate through baryonic shells. The clump shells are assigned some angular momentum (similarly to the baryon and dark matter shells) that stops them from reaching the singularity at the centre. Clump shells typically
oscillate around the halo’s centre before the processes described below destroy them.

Each shell is assumed to contain \( n_{\text{cl}} \) clumps with mass

\[
m_{i} = \frac{M_{\text{shell}}}{n_{\text{cl}}}. \tag{2}
\]

\( M_{\text{shell}} \) and \( m_{i} \) are the total shell mass and the mass of each gaseous clump, respectively. The shells interact with the baryons by decelerating according to equation (1). The drag force equation and energy equation of interaction between some clump \( i \) and a parcel of gas are

\[
f_{\text{cl}}^{i} = -F_{\text{gas}}^{i} \tag{3}
\]

and

\[
E^{i} = f_{\text{cl}}^{i} v_{\text{cl}}^{i} + F_{\text{gas}}^{i} u_{\text{gas}} + Q^{i} = 0, \tag{4}
\]

respectively, with \( f_{\text{cl}}^{i} \) and \( F_{\text{gas}}^{i} \) the forces on the clump and gas parcels arising from the clump–gas interactions, and \( v_{\text{cl}}^{i} \) and \( u_{\text{gas}} \) the velocities of clump \( i \) and the gas, respectively. \( Q^{i} \) is the rate at which clump \( i \) heats the gas parcel it is embedded in at that time. We assume that many clumps are present within each parcel of gas, and that their motions are isotropic so their forces cancel out:

\[
\sum f_{\text{gas}}^{i} = 0 \tag{5}
\]

and

\[
E_{\text{tot}} = \sum f_{\text{cl}}^{i} v_{\text{cl}}^{i} + \sum F_{\text{gas}}^{i} u_{\text{gas}} + \sum Q^{i} = \sum f_{\text{cl}}^{i} v_{\text{cl}}^{i} + \sum Q^{i} = 0. \tag{6}
\]

In the reference frame of a gas parcel, the clump loses energy, which is converted to heat, so using equation (1) we identify

\[
Q^{i} = f_{\text{drag}} v_{\text{rel}} = \frac{1}{2} C_{A} \rho_{\text{b}} |v_{\text{rel}}|^{3} \tag{7}
\]

as the heating rate that clump \( i \) heats the gas parcel in which it is embedded. The total energy of the clumps and of the gas should be conserved on average according to equation (6), assuming there are enough clumps so the averaging is correct and that the assumption that the clumps have isotropic velocities is good. The difference equations in HYDRA conserve energy algebraically\(^2\) (making energy conservation independent of resolution), and we do not want to violate this property. We thus require a detailed balance between the clump deceleration and energy deposition of each clump:

\[
f_{\text{cl}}^{i} v_{\text{cl}}^{i} = Q^{i}. \tag{8}
\]

which implies

\[
f_{\text{cl}}^{i} = f_{\text{drag}} \frac{v_{\text{rel}}}{v_{\text{cl}}}
\]

\[
= \frac{1}{2} C_{A} \rho_{\text{hot}} v_{\text{rel}}^{2} \frac{|v_{\text{rel}}|}{v_{\text{cl}}} = \frac{1}{2} C_{A} \rho_{\text{hot}} \left| \frac{v_{\text{rel}}}{v_{c}} \right|^{3}. \tag{9}
\]

Equation (9) recovers the exact solution when the gas parcel is at rest (which is the typical case of heating of a hydrostatic gas) and when there is no relative velocity between the gas and the clump (i.e. no drag).

\(^2\)An algebraic scheme is such that the exact energy term is always added to one component and subtracted from the other explicitly, making sure the energy is conserved to the machine accuracy.

Physically, the drag forces always act to decrease the radial and tangential components of the velocity of clumps. While the radial velocity is replenished by the gravitational force, the tangential velocity monotonically decreases in time, so clumps lose angular momentum, becoming more radial in their trajectories. Since the angular momentum of the clumps shells in our simulations is conserved, the clumps in the simulations cannot spiral towards the centre. To compensate for this problem, and allow clumps to reach the central core, a much smaller angular momentum is assigned to the clumps, placing them on almost radial trajectories.

2.1.3 Fragmentation of clumps

Once the framework for accelerations and heating is defined, we proceed to implement recipes for clump evolution. In this work, we implemented only the most crucial additional recipes from DB08: clump fragmentation and clump destruction. A clump fragments into \( n_{\text{frag}} \) clumps (two throughout this work) once it has repelled its own mass of ambient gas. The amount of gas expelled is calculated by numerically integrating over \( \pi r_{\text{cl}}^{2} \rho_{\text{gas}} v_{\text{rel}} \, dt \) in the same Runge–Kutta scheme used for the dark matter. After each time-step, the column mass and clump mass is compared. When the column mass exceeds \( m_{\text{cl}} \), \( m_{\text{cl}} \) is divided by \( n_{\text{frag}} \), \( n_{\text{cl}} \) is multiplied by \( n_{\text{frag}} \) and the column mass integral is reset to 0. As \( m_{\text{cl}} \) becomes exceedingly small, its drag deceleration becomes large, and the periods between consequent fragmentation events become shorter. Clumps are destroyed when their mass decreases below a critical mass, at which conduction and evaporation is expected to disintegrate them completely (in this work – \( 10^{4} M_{\odot} \); DB08).

2.1.4 Destruction of clumps

The destruction of clump is achieved by adding its mass to the corresponding baryonic shell. The velocity and angular momentum are not changed, and the internal energy and temperature are calculated by mixing the cold and hot components according to energy conservation by solving for the final internal energy, \( E_{\text{int}}^{i} \), in

\[
M (E_{\text{int}} + E_{\text{kin}} + E_{\text{grav}}) + m (e_{\text{in}} + e_{\text{kin}} + e_{\text{grav}})
\]

\[
= (M + m) (E_{\text{kin}} + E_{\text{grav}}) + (M + m) E_{\text{int}}^{i}. \tag{10}
\]

with \( M, E \) correspond to the baryonic shell values, \( m, e \) to the clump shell. All energies are specific energies (per unit mass) and \( e_{\text{int}} = C_{e} T_{\text{cl}} \) with \( T_{\text{cl}} = 10^{4} \text{ K} \). Rarely, the final temperature will drop below \( 10^{4} \text{ K} \), at which case the final temperature is set to \( 10^{4} \text{ K} \). This is found to occur only for baryonic shells that are cold (around \( 2 \times 10^{4} \text{ K} \)) and on a free fall to the BCG, and the temperature floor that is artificially applied never stops their infall. In this case, the fictitious energy is tracked throughout the run and is always negligible. The overall accreted mass is not effected by this correction.

2.1.5 Applicability to 3D simulations

In the 1D case described above, the gas parcel is a finite-dimension gas shell, and the clump is a thin shell. While 3D simulations are beyond the scope of this work, if one wishes to incorporate the effects of cold clumps in 3D simulations, a generalization for the 3D case is readily available, as follows. Both SPH and grid-based (Eulerian or Lagrangian) simulations that model dark matter as \( N \)-body particles can assign clumps to a dark matter particle according to equation (2), calculate its acceleration according to equation (9) and heat the gas according to equation (7). In cosmological simulations, it is also necessary to self-consistently create those clumps.
These can either be created semi-analytically according to cooling instabilities (Field 1965; Binney, Nipoti & Fraternali 2009) or by identifying unresolved gaseous clouds and replacing them with clump particles (Kereš & Hernquist 2009 identify clump formation on Milky Way (MW) scales, but resolution would not allow to scale this procedure to galaxy cluster scales).

2.2 Cell splitting – a 1D adaptive mesh refinement

In the Lagrangian formalism, the amount of baryonic mass within each shell is constant throughout a simulation. Clump destruction, however, violates that by dumping mass into the baryonic shells at the event of clump destruction. This mass deposition is expected to occur preferentially at specific regions. Indeed, the simulations discussed below initially caused the formation of a few extremely massive shells that absorbed most of the mass of the clumps. This situation (of large contrasts between adjacent shells in mass and widths) reduces the accuracy of the numerical scheme, and effects the accretion rates on to the BCG. To overcome this, an adaptive splitting of baryonic shells is performed: when the shell is wider by some factor than both its neighbours, or when the shell’s width divided by its radius exceeds some preset fraction, the shell is split to two. The factors ultimately used were \( \Delta r > 4 \max(\Delta r_{n-1}, \Delta r_{n+1}) \) and \( \Delta r_{n}/r_n > 0.2 \) (with \( \Delta r_n \) and \( r_n \) the width, and central radius of shell \( n \)) which caused splitting to occur a few dozens times throughout a full, Hubble time simulation. This choice of parameters is a result of trial and error motivated by the requirement that except for transient periods, the width of Lagrangian shells is roughly equally spaced (locally), ensuring the robustness of the numerical scheme.

A shell is split into two constant mass shells. Values that are naturally defined at centres of shells (density, temperature) are treated as step functions and their values are equal in the two new shells. Values that are defined on boundaries (radius, velocity, angular momentum) are interpolated linearly with mass. This definition ensures that the total internal energy of the system remains the same. The potential and kinetic energies, however, can change. This energy non-conservation is not corrected for explicitly. Rather, it is treated as non-conservation and tracked throughout the run. In the high-resolution simulations shown below (2000 baryonic shells and 10000 clump and dark matter shells), the total, overall energy is conserved to better than \( 10^{-2} \) over a Hubble time.

2.3 Convection and mixing length theories

2.3.1 Convection

A long-term balance between cooling and heating requires, in addition to sufficiently large energy injection, a correct distribution of this energy. The cooling rate, at constant pressure, scales as \( \epsilon_{\text{cool}} \sim \rho^{0.5} \), and most heating mechanisms (the drag forces discussed here, dynamical friction, radiative heating from supernovae and AGN) generally heat according to \( \epsilon_{\text{heat}} \sim \rho \). Even if at some point in space, and at some initial time, \( \epsilon_{\text{cool}}/\epsilon_{\text{heat}} = 1 \), this ratio scales like \( \rho^{0.5} \). This relation indicates a positive feedback so if density increases, cooling is more efficient, causing a further increase in density at constant pressure, and unstable cooling will occur. Vice versa, a density decrease increases the relative importance of heating over cooling, decreasing the density further, and an overheating instability occurs. This point has been made by Croton & Ostriker (2008) (see, however, Kunz et al. 2011, who claim that residual magnetic turbulence might inject energy into the gas in a stable manner), using 1D hydrodynamic calculations of clusters initially hydrostatic within a static potential well. They find that stable, long-term equilibrium requires fine-tuning of the heating efficiency which is unlikely. The present work differs in that it treats the gravitational drag feedback and the cluster evolution from initial cosmological perturbation consistently, but should still suffer from a heating instability. Such an overheating will manifest as a shell or region of shells of gas continuously becoming hotter and underdense, with very high entropy. In a 3D configuration, this entropy inversion is unstable to convection when entropy is declining outwards: a slightly underdense parcel of gas floats buoyantly, carrying energy and momentum outwards, and overdense parts sink towards the centre reducing the average entropy and specific energy of the core. For our 1D simulation we invoke a 1D subgrid model for convection–mixing length theory (Spiegel 1963).

2.3.2 Mixing length theory

Convection occurs when an adiabatic displacement of a parcel of gas, in pressure equilibrium with its new position, results in a net force on that parcel tending to increase the displacement. This requires that, for an upward displacement, the temperature of the gas parcel will be smaller than its surrounding:

\[
\Delta V T = \left( \frac{\partial T}{\partial r} \right)_s - \left( \frac{\partial T}{\partial r} \right)_{S} \]

\[
= \left[ \left( \frac{\partial T}{\partial S} \right) \frac{\partial S}{\partial r} + \left( \frac{\partial T}{\partial P} \right) \frac{\partial P}{\partial r} \right] \left( \frac{\partial T}{\partial P} \right) - \left( \frac{\partial T}{\partial P} \right) \left( \frac{\partial T}{\partial P} \right) \left( \frac{\partial P}{\partial r} \right) \]

\[
= \left( \frac{\partial T}{\partial S} \right) \frac{\partial S}{\partial r},
\]

with the first term in the right-hand side (rhs) corresponding to the actual temperature derivative in the profile, and the second to the adiabatic change in temperature as a result of the pressure profile of the halo. A negative value of \( \Delta V T \) allows for convection to occur. It is convenient to derive this relation using a form of the ideal equation of state in which the two thermodynamic free parameters are the entropy, \( S \), and the pressure, \( P \):

\[
P = \frac{N_A k_B}{\mu} \rho T,
\]

\[
S = \frac{N_A k_B}{\mu} \ln \frac{T^{3/2}}{\rho},
\]

(12)

can be inverted (setting \( \bar{\mu} = \mu/N_A k_B \)) into

\[
\rho = (\bar{\mu} P)^{1/3} e^{-(2/5) \bar{\mu} S},
\]

\[
T = (\bar{\mu} P)^{-1/3} e^{(2/5) \bar{\mu} S}.
\]

(13)

Plugging these relation into the derivatives in equation (11) we get

\[
\Delta V T = \frac{2}{\bar{\mu}} T \frac{\partial S}{\partial r},
\]

(14)

recovering the known results that when the composition of the gas is constant, entropy inversion leads to convection.

The buoyant bubbles rise for a typical length before being destroyed by Kelvin–Helmholtz instability or conduction. The details of this destruction depend on the size of the bubbles, the smoothness of the density and gravitational profile and conduction and magnetic fields. Solving for it requires fine 3D simulation of the convective process. We replace this dependency by a free dimensionless parameter, the mixing length (L), assuming that bubbles rise a distance,
Gravitational heating in clusters by clumps

2.4 The observed properties of hot gas with cold clumps

In the current implementation, when clumps are destroyed, their mass is added to the hot component instantly. In reality, the process of Kelvin–Helmholz (KH) fragmentation, followed by small-scale evaporation and conduction of the debris will yield a multiphase gas, with an effective entropy and temperature which is between the values of the hot and cold phases. As will be shown later (Section 4), the clump density and clump destruction rate increase towards the centre so effective cooler and lower entropy values are expected there. The radiative signature of gas heating through all the temperatures between $10^6$ K to the cluster ambient gas temperature of $\sim 3 \times 10^7$ K is expected to be significantly different from that of radiative cooling since it is governed by heating processes (emission spectrum from heating gas, albeit by other heating mechanisms have been studies in Voit & Donahue 1997; Oh 2004). A framework of heating and cooling processes in layers between hot and cold media have been proposed by Begelman & Fabian (1990) and Gnat, Sternberg & McKee (2010).

The observational signature of clump breakup would require detailed 3D simulation of clump interactions with cluster core gas, and multiphased modelling of the radiative signature during the heating process, and is beyond the scope of this work. Instead, we will plot below the mass weighted entropy and temperature of the two components. This is a lower limit for the observed entropy and temperature as the clumps’ contribution to the luminosity, particularly at X-ray wavelength, is probably small. Physically, it corresponds to the thermodynamic properties expected in the event of full mixing between the cold and hot phase. The actual temperature and profile expected from the multiphase gas is thus bracketed between the hot only component, and the mass weighting between the hot and cold components presented in Figs 3 and 4.

3 RESULTS

The simplified model described in Section 2 spans a multidimensional parameter space including the fraction of accreted gas in clumps, the initial clump masses, the number of fragments that a clump breaks up to and the mixing length parameter for convection. In the absence of additional physical insight concerning the formation mechanism and properties of these clumps, we attempt to find a working set of values for the model parameters, to serve as a feasibility test and hopefully provide clues for acceptable values of the key parameters. Ultimately, a more systematic survey of parameter space will have to be conducted, with physically motivated values for key parameters such as number and mass of clumps.

\[ H_p = L \frac{P}{\rho g}, \]

with $g = GMl^2r$. A typical acceleration (assuming isobaric perturbations) is

\[ a = \frac{\delta \rho}{\rho} = \frac{g}{\rho} \left( \frac{\partial \rho}{\partial S} \right) \frac{\partial S}{\partial r} H_p, \]

so a typical bubble velocity is

\[ v = (2aH_p)^{1/2} = H_p \left[ 2 \frac{g}{\rho} \left( \frac{\partial \rho}{\partial S} \right) \frac{\partial S}{\partial r} \right]^{1/2}. \]

Once the velocity exceeds the local sound of speed in the halo, shocks are created which quickly act to mix the bubble with its surrounding. In the following calculation the velocity is not allowed to exceed the speed of sound, $c_s$, and in that case $H_p$ is reduced until $v = c_s$ in equation (17). A maximal convection model is a model with arbitrary high $L$, so effectively the bubbles always accelerate until the speed of sound, at which time they are broken and mixed.

The flux of energy per unit surface per unit mass is

\[ F_e = C_p v [\Delta \nabla T] H_p = C_p v \left[ \frac{2}{5} \frac{\mu}{\rho} \frac{\partial S}{\partial r} H_p \right], \]

which is determined by the halo profile from the simulation at each time, and the mixing length $L$. $C_p$ is the constant pressure heat capacity and is related to $\mu$ by $C_p = 5/(2\mu)$.

A numerical solution of the mixing length model requires evaluation of the incoming and outgoing fluxes from the boundaries of each radial shell. The fluxes depend on the temperature gradient between each shell and the ones directly below and above it, and interpolation of thermodynamic properties from the shell centres to the shell’s edges is required. A solution using an explicit numerical scheme (with the fluxes determined at the beginning of each time-step) requires extremely small time-steps to avoid negative temperatures between 10$^6$ K and the spatial resolution limit (thin red line of Fig. 3). These extreme gradients (that are also present at edges of radio bubbles in clusters) are far from linear perturbations, and the validity of the linear approximation of the various convection prescriptions is highly questionable.

3 See, for example, the short-dashed lines in figs 6 and 10 of Parrish et al. (2008).
In this section we restrict ourselves to one typical CC cluster halo with virial mass of $3 \times 10^{14} \, M_\odot$ by $z = 0$, a diffuse baryon fraction of 10 per cent and a smooth accretion history according to the average growth rate of the main progenitor á la Neistein et al. 2006 (see Birnboim et al. 2007 for a detailed description). The metallicity is assumed to be constant at $Z = 0.3 \, Z_\odot$, and the cooling is bremsstrahlung and line cooling according to Sutherland & Dopita (1993). The initial resolution is 2000 baryonic shells and 10000 dark matter shells, roughly logarithmically spaced in their initial radii. When shells expand, the adaptive mesh refinement algorithm splits them (Section 2.2), so the resolution near the cluster core ($\sim 50$ kpc) at all times is better than $\sim 2$ kpc. This yields converged results in terms of the profiles and the amount of gas that cools. We implement three different models for clump heating, as follows.

(i) Model C is the null model with no clump heating and no convection. It is meant to reproduce the overcooling problem.

(ii) Model CH adds clump heating but no convection, so it should show the overheating instability. The fraction of baryons in clumps is 5 per cent and the clump initial mass is $10^8 \, M_\odot$. The clumps are simulated by 10000 clump shells (Section 2.1.2).

(iii) Model CHC has the same clump heating as in CH but with maximum convection turned on (Section 2.3.2). The smoothing by convection is supposed to regulate the clump heating and yield relaxed clusters compatible with observations.

Fig. 1 shows the time evolution of the gas in our simulated cluster comparing the three different models for cooling, clump heating and convection. The initial Hubble expansion and consequent turnaround of the Lagrangian gas shells is clearly seen, and the virial shock can be easily identified after a collapse by a factor of $\sim 2$, both by a jump in temperature from below $10^4$ to above $10^7 \, K$, and by the abrupt slow down of the infall velocity, which is almost brought to a halt behind the shock. The global large-scale properties of the cluster are not affected by the addition of heating and convection. The virial radius evolves in a similar way, and the typical temperature in the halo at $z = 0$ remains at $T \sim 2 - 3 \times 10^7 \, K$ ($\sim 2 - 3$ keV), consistent with the expected virial temperature of $2.2 \times 10^7 \, K$ for a cluster of virial mass $3 \times 10^{14} \, M_\odot$. However, the models differ at the core, within the innermost 100 kpc especially during the last 6 Gyr of evolution. Model C shows inward cooling flows at all times, as expected (Fabian 1994). With the addition of clump heating in model CH, the cooling flows are stopped before $t = -6$ Gyr, the gas in certain shells is overheated to extreme temperatures $\gtrsim 10^8 \, K$, these shells are interlaced with cooler shells of $\sim 10^6 \, K$ and together the whole core inflates. This behaviour is in conflict with the relaxed nature and smooth entropy and temperature profiles of CC clusters (Donahue et al. 2006). The addition of
Gravitational heating in clusters by clumps

2573

Figure 2. Time evolution of X-ray luminosity (top), BCG mass (middle) and BCG accretion rate (bottom), all smoothed over 1 Gyr.

convection in model CHC removes the local overheating, and keeps the core in equilibrium at the virial temperature with no cooling flow. A more detailed comparison of the models and observations follows.

3.1 Time evolution of X-ray luminosities and BCG masses

Fig. 2 shows the evolution of X-ray luminosity, the mass of the BCG, and the accretion rate on to it, for the three different models. The BCG is represented by the mass in the ‘disc’, the mass that is supported by angular momentum in the 1D simulations, extending to \( \sim 10 \) kpc. The X-ray luminosity is obtained as the total cooling radiation from the gas outside the BCG.\(^4\) The quantities are smoothed over 1 Gyr to erase sharp features that result from the discreteness of the calculation in the idealized spherical calculation.

Model C shows a variable luminosity, which occasionally exceeds by an order of magnitude or more the observed luminosity of \( \sim 10^{44} \) erg s\(^{-1}\) as derived from the \( L_X-T \) relation (Edge et al. 1990; David et al. 1993; Markevitch 1998). In model C the final BCG mass exceeds \( 2 \times 10^{12} \) M\(_\odot\) and the accretion rate, which is an indicator for the star formation rate, has long episodes where it is in the range \( 100-1000 \) M\(_\odot\) yr\(^{-1}\), at odds with observed CC clusters. When clump heating is added in model CH, the cooling flow into the BCG is drastically suppressed since before \( t \sim -8 \) Gyr, and it reaches a complete shutdown after \( t \sim -5 \) Gyr. The luminosity maintains a high level of \( \sim 2 \times 10^{45} \) erg s\(^{-1}\) since \( t \sim -7 \) Gyr. This results from the overheating instability in the halo core, leading to very dense shells that boost the dissipation due to drag interaction with the clumps and enhance the resulting radiation. The addition of convection in model CHC brings the BCG mass to a constant value of \( \sim 10^{12} \) M\(_\odot\) with no detectable cooling flow since \( t \sim -6 \) Gyr. The smoothing of the instability brings the luminosity to a low value of \( \sim 10^{44} \) erg s\(^{-1}\), consistent with the observed \( L_X-T \) relation.

3.2 Entropy and temperature profiles

The \( z = 0 \) entropy profiles of the three models are plotted in Fig. 3 and the temperature profile of model CHC is plotted in Fig. 4. The entropy profiles in the outer halo, overall increasing close to linearly with radius and having values \( \sim 100 \) keV cm\(^2\) at 100 kpc, are similar in all models and consistent with observations of CC clusters (Donahue et al. 2006; Cavagnolo et al. 2009). The apparent periodic fluctuations represent cold fronts that result from mergers of outward-propagating shocks at regular intervals with the virial shock (also visible in Fig. 1), which have been studied in Birnboim, Keshet & Hernquist (2011). While the fluctuations in models C and CH are locally non-monotonic, the entropy profile of model CHC is monotonically increasing throughout because the convection removes negative entropy gradients. Model CH shows a high-entropy core due to overheating, and a strong variability representing a mixture of cold-dense and hot-dilute shells that result from the overheating instability, both in conflict with observations. The convection introduced in model CHC removes the fluctuations and produces what seems to be a flat core entropy profile.

\(^4\) Most of the energy is emitted from cooling of \( \sim \) keV gas, and is predicted to contribute to the X-ray luminosity that is relevant to the \( L_X-T \) relation (e.g. Markevitch 1998).
inside 150 kpc at 50 keV cm$^2$. Such a core is still inconsistent with CC cluster cores, but recall that the profile shown is limited to the entropy of the hot component alone. The effective entropy that could actually be observed is a mixture of the entropy in the hot gas, cold clumps, and anything in between as discussed in Section 2.4. The effective entropy profile shown in Fig. 3 is monotonically increasing in the core down to $\sim$30 kpc, consistent with CC clusters.

The temperature profile of model CHC at $z = 0$ is plotted in Fig. 4. It shows a roughly isothermal halo at the virial temperature outside the core of $\sim$100 kpc, with a mild decline toward the virial radius, as observed (Donahue et al. 2006). This large-scale temperature profile has not been affected much by clump heating and convection. The temperature of the hot component is rising toward the centre, by a factor of $\sim$2, but the effective mass-weighted temperature of the mixture of hot and cold components is declining toward the centre, by a factor of a few. This mass-weighted effective temperature (that corresponds to the single temperature the gas would if the phases were fully mixed) is a lower limit to the luminosity-weighted temperatures observed, and is consistent with the moderate temperature decline in CC cluster cores. The temperature profiles of the various models, and its time evolution, can also be seen in Fig. 1.

### 3.3 Sensitivity to choice of parameters

The results presented in this section indicate that clumps of $m_c = 10^7 M_\odot$ that make for $f_c = 0.05$ of the gaseous component of clusters is sufficient to remedy the overcooling problem, and quench the accretion of gas on to the BCG. These parameters were picked to demonstrate the effectiveness of our model. We find that our simulated clusters are not particularly sensitive to these values, and that no fine-tuning is required. Rather, a wide envelope of allowed parameters is allowed. Fig. 5 is analogue to Fig. 2 and compares the luminosity, BCG mass and accretion rate predicted by simulations with maximal convection for a parameter choice of ($m_c = 10^7 M_\odot, f_c = 0.1$) and ($m_c = 10^8 M_\odot, f_c = 0.1$), along with our fiducial model of ($m_c = 10^8 M_\odot, f_c = 0.05$). The same behaviour is also found in the resulting radial profiles. A simulation with the parameters ($m_c = 10^7 M_\odot, f_c = 0.05$) allowed for too much cooling, and final BCG mass of $1.5 \times 10^{12} M_\odot$ – that is slightly excessive.

Beside these parameters, some of the theoretical model assumptions (for example the drag efficiency, the fragmentation of clumps and the convection) might need to be modified after more detailed 3D simulations or additional observational constraints are found. We expect that for a different model the parameters will need to be readjusted, but since the heating model is not particularly sensitive, we expect that such a choice will always be possible. We do not believe that a comprehensive parameter survey will be beneficial at this point, until the different components of this model are better constraint either theoretically or observationally.

### 4 OBSERVATIONAL SIGNATURES OF CLUMP HEATING

The model parameters chosen in this paper were motivated by the analysis of DB08 and were calibrated such that the model crudely reproduces the properties of CC clusters in order to demonstrate the feasibility of such a model of gravitational heating. The reproduced properties include the BCG mass, the cold mass accretion history, the X-ray luminosity and the entropy and temperature profiles. This
model makes additional predictions that could distinguish it from other heating models such as the ones based on AGN feedback. Some of these predictions are discussed here.

4.1 Cold gas in the ICM

The mass function of clumps at every radius is a distinctive prediction of our proposed model. Assuming an initially uniform population of $10^8 \, M_\odot$ clumps, Fig. 6 shows the profiles of number density and average clump mass for models CH and CHC. The fraction of volume occupied by cold clumps is shown in the bottom panel. These predictions depend on the specific choice of the initial mass of the clumps and on the baryonic fraction of mass in these clumps. The number density of clumps increases toward the centre $\propto r^{-2}$, partly reflecting the general density profile of the cluster and partly because of the breakup of clumps into fragments (Section 2.1). Once clumps are fully destroyed (fragments $< 10^4 \, M_\odot$), their mass is added to the hot component, and they are no longer plotted in Fig. 6. The figure thus shows the steady state population of clumps as they are continuously accreted and destroyed. The average clump mass is declining from the initial value of $10^8 \, M_\odot$ in the outer halo to $\sim 10^5 \, M_\odot$ near the centre, reflecting the clump fragmentation as they flow in. The fraction of volume occupied by the cold clumps peaks at a few per cent near the cluster core, and drops to smaller values at larger radii. This justifies ignoring the additional pressure caused by this component.

The clumps are initially cold, and, except for the mild compression they undergo as they fall in following the increasing pressure of the ambient gas, they are not expected to heat up or emit much radiation. However, as the clumps are disrupted by hydrodynamical instabilities and possibly also by tidal effects, they fragment into smaller pieces for which conduction and evaporation become more important (Section 2.4). Once heated to intermediate temperatures, the gas begins to radiate. Spectroscopic observations could in principle constrain the validity of this model in comparison with AGN-feedback models, where one expects gas cooling rather than heating through the intermediate temperatures. The current simplified implementation of clump heating does not permit a proper comparison, which is left for future work. Additionally, three-dimensional simulations are required for a detailed analysis of the shape of the clumps as they are stretched perhaps leading to morphologies resembling filaments (Murray & Lin 2004). This emission, in Hα and line and continuum emission of the intermittent X-ray temperature gas, may allow more accurate comparisons of this model with the observed profiles in cluster cores. Observations in the Perseus cluster (NGC 1275; Conselice et al. 2001; Fabian et al. 2008) show a complicated structure of Hα filaments and blobs. The typical masses of these features are $10^4–10^5 \, M_\odot$, consistent with the allowed mass range for clumps in DB08, and with the distribution predicted by our model (Fig. 6). We note that this result depends on the initial mass of the clumps – a free parameter here. The consistency of this prediction with observation is an indication that our choice of initial mass of $10^8 \, M_\odot$ is reasonable. Fabian et al. (2008) invoked strong magnetic fields to stabilize the filaments for cosmological times, such that their age can match that of the observed radio bubbles. Our heating model suggests instead that these filaments are constantly being destroyed, as new clumps enter the cluster core, get stretched and destroyed, and create new filaments. The projected filling factor of these structures approaches unity within the innermost 10 kpc and it drops outwards (Conselice et al. 2001). Such a behaviour is predicted by our model (Fig. 6). Hα emission in other clusters have been reported by Heckman et al. (1989), who found that the cold gas has velocities at random directions rather than a coherent radial cooling-flow pattern. This kinematics could be interpreted as clumps oscillating in and out at the vicinity of the BCG. Structures of neutral gas are also seen in the Virgo cluster by the Arcicio Legacy Fast ALFA (ALFALFA) 21-cm survey (Giovanelli et al. 2007; Kent et al. 2007) showing evidence for neutral gas arranged in clumps, with masses as low as the detection limit of $2 \times 10^5 \, M_\odot$, sometimes with no optical counterparts.

4.2 Turbulence in the ICM

Another prediction of this model is the power of turbulence that is produced in the ICM. When the clump velocities are subsonic with respect to the ICM, the energy and momentum of the drag are first converted to kinetic energy in turbulence, which cascades down to smaller scalelengths where it dissipates into heat. When the motion is supersonic, some of the energy is converted directly into heat, but since momentum is conserved, some of the energy must be transferred as kinetic energy to the ICM. We assume that the turbulence is generated on a scalelength comparable to the average distance between clumps, $L$, and that a Kolmogorov spectrum governs the cascade of eddies from this scale to smaller scales. The turbulent energy and pressure are then

![Figure 6. Average number density of clumps (top), average clump mass (middle) and fractional volume in clumps (bottom) as a function of radius at $z = 0$ for models CH and CHC.](https://academic.oup.com/mnras/article-abstract/415/3/2566/105022/1522581/105022/guest)
4.3 High-velocity clouds in the galactic halo

High velocity clouds (HVCs) are observed in 21-cm H$\alpha$ data (Blitz et al. 1999) as concentrations of gas moving at velocities $>100$ km s$^{-1}$ relative to the rotating frame of the MW. Options for the spatial origin of the HVCs range from the MW disc (Wakker et al. 2008, and references therein), through the Magellanic Clouds (Olano 2008), to extragalactic origin (Blitz et al. 1999). The distance to and the ionization fraction of individual clouds (and therefore their size and mass) are unknown. Putman et al. (2003) estimated distances of HVCs based on their He$\alpha$ flux and models for the emission of ionizing radiation from the MW. They find, within the modelling and measurement uncertainties, that most HVCs are within a distance of $\lesssim 30$ kpc, indicating a mass range of $10^4$–$10^6$ M$_\odot$ (Putman et al. 2003; Birnboim & Loeb 2009). Other estimates by absorption features (Thom et al. 2008) yield comparable distances and masses. The origin of the HVCs is unclear. Models have proposed that they form within the Galactic halo by cooling instabilities (Maller & Bullock 2004; Kereš & Hernquist 2009). See, however, Fraternali & Binney (2008), Binney et al. (2009) analysed formation of clumps in smooth MW halo conditions from thermal instability and concluded that it is unlikely to occur near the centre. They find, however, that the conditions become favourable closer to the halo virial radius and when the entropy profile is shallow (as is shown numerically in Kaufmann et al. 2009). Their stability analysis tests growth of instability from infinitesimal perturbations but the growth of non-linear perturbations caused by shocks, collisions and gravitational perturbers depends on initial conditions. The line-emission peak of the cooling curve at $\sim 10^8$ K would make the warm cosmic filaments outside clusters more susceptible for clump formation. However, clumps have been shown to be a natural consequence of cold-flow filament breakup by hydrodynamic instabilities (Kereš & Hernquist 2009). For example, streams that do not flow radially to the halo centre are susceptible to Rayleigh–Taylor instability. Furthermore, shocks that originate from the galaxy, e.g. by mergers or by starbursts, are likely to form clumps by Richtmyer–Meshkov instability. The destruction of these clumps, and their interaction with the Galaxy and the intergalactic medium, are likewise under debate (Fraternali & Binney 2008; Kereš & Hernquist 2009).

We point at an obvious analogy between the observed Galactic HVCs and the clumps addressed in this paper. Their masses are comparable, and their spatial distributions in the halo and toward its centres are possibly similar. The larger pressure in the ICM of a more massive halo would make the cluster clumps denser than the Galactic HVCs (by the ratio of virial temperatures which is more than 10 times larger for clusters), so clumps of a similar mass could survive longer in clusters, allowing them to travel to the centre according to the estimates in Section 2. The total mass encompassed in HVCs seems to be $\gtrsim 10^8$ M$_\odot$, making it a few per cent of the total baryons in the MW halo, in good agreement with our fiducial choice of parameters. A missing piece of the model is the yet unspecified origin of the clumps, in terms of physical mechanism and location. The existence of HVCs provides circumstantial evidence that such clumps might form. As long as the clumps are formed before they fall into the halo, or even if they form inside the halo at a radius that is not much smaller than the virial radius (Maller & Bullock 2004; Kereš & Hernquist 2009), the gravitational energy that is released during their infall is significantly larger than the energy required for heating the clumps to the virial temperature (DB08).

5 DISCUSSION AND CONCLUSION

The concept of gravitational heating of ICM gas as a partial or full solution to the cooling-flow problem is more general than the specific clump model discussed in this paper. It is easy to show that the gravitational energy that is released as baryonic matter falls in through the halo potential well is enough to balance the cooling rates in groups within haloes of virial masses $\sim 10^{13}$ M$_\odot$, and it exceeds the cooling rate by more than an order of magnitude in...
cluster haloes $\sim 10^{14}$–$10^{15}$ M$_\odot$ (DB08). This point has also been made in Fabian (2003), Wang & Abel (2008) and Khochfar & Ostriker (2008). El-Zant et al. (2004), Faltenbacher & Mathews (2007) and Khochfar & Ostriker (2008) tap into the same energy source, but utilize dynamical friction that is less effective. Naab et al. (2007) and Johansson et al. (2009) show, however, that dynamical friction can efficiently stop gas accretion on to massive elliptical galaxies. Conduction also taps into this energy source, and was proposed three decades ago (Bertschinger & Melikson 1986; Rosner & Tucker 1989). It is probably ruled out because of magnetic field suppression of the conduction (Binney & Cowie 1981; Fabian 1994, and reference within), and because the resulting profiles would have a flat temperature core (Bregman & David 1988). See, however, Narayan & Medvedev (2001) and Kim & Narayan (2003) for alternative ideas. Here we use the same energy source, but the physical process used to couple the energy with the baryonic cooling component is hydrodynamic drag for which the strength of interaction peaks at low clump masses and high velocities.

The challenge for every heating model is to distribute the energy uniformly throughout the cluster core both in space and time (De Young et al. 2008; Cattaneo et al. 2009). In order to obey the observational constraints, the heating mechanism should suppress the gas mass that actually cools by two orders of magnitude. Such a shutdown requires that the mechanism should act smoothly over a scale of a few kpc, set by the smallest object that would cool in the absence of feedback while being continuously heated by conduction from its surrounding. This scale follows from

$$L_{\text{cool}} \sim \sqrt{\eta k_s P} = 7 \eta_0 T_2^{5/2} n_e^{-1} t_{\text{cool}} \text{ kpc}, \quad (20)$$

with $k_s$, the Spitzer coefficient (Spitzer 1962), $\eta$ the reduction of the Spitzer coefficient due to magnetic fields, $t$ the cooling time of the gas and $T_2 = kT/2\text{keV}, n_e = n/10^{-2} \text{cm}^{-3}, t_{\text{cool}} = t/10^8 \text{yr}, \eta_0 = \eta/0.2$. The value $\eta = 0.2$ is a reasonable upper limit for the efficiency of conduction (Narayan & Medvedev 2001). It also has to be temporally smooth over the cooling time-scale, which is a few $10^8$ yr at most (Donahue et al. 2006). The fact that the energy source of gravitational infall is automatically distributed over the cluster volume and over Hubble times makes it easier to meet this challenge with gravitational heating than it is with AGN-feedback models, where the source is on scales 10 orders of magnitude smaller than the cluster scales. However, the coupling of the infalling baryons with the ICM should be such that most of the gravitational energy is deposited in the cluster core.

This paper addressed one specific scenario of gravitational heating, in which the mechanism for feeding the energy into the ambient ICM is via hydrodynamic drag acting on small clumps of cold gas. In a typical cluster, with a modest gas fraction $\sim 5$ per cent of the accreted baryons in cold clumps of $\sim 10^6$ M$_\odot$, the clump heating suppresses the cooling flows toward the BCG for the last 6 Gyr. The conditions at the core do not affect the incoming clumps, so the process is not strictly self-regulating. Regardless, we find that the large-scale properties of the cluster, such as the overall gas fraction, virial shock and temperature outside the core, are unaffected by the heating. Furthermore, the core does not explode: it reacts to the heating smoothly and quiescently without a need for inherent self-regulation. Because of the overheating, the central density decreases and the total X-ray luminosity emitted from the core (Fig. 2) declines such that the core obeys the observed $L_X\sim T$ relation. On cluster core scales, convection acts to flatten the entropy profile of the hot component and carry heat outwards as expected. The effective entropy profile after taking into account the cold clumps as well does not exhibit an entropy core, and is consistent with observed entropy profiles, to within the model limitations that are discussed. A local instability caused by the linear dependence of the heating rate on density ($\rho_{\text{heating}} \sim \rho$) acts to create extreme entropy and temperature peaks of sub-kpc scales. These peaks create strong convection that, once accounted for in the simulations, stabilize the heating process on local scales as well.

With the fiducial values of the parameters used in this work in a $3 \times 10^{14}$ M$_\odot$ cluster halo, the model is successful in quenching the cooling flows and in reproducing adequate BCGs, X-ray luminosities and entropy and temperature profiles. The model also predicts the expected level of turbulence in clusters, and the fraction of cold gas as a function of radius. These two observables are predictions of this model, while they are not naturally addressed by AGN-feedback models.

In DB08 we analysed clump heating in a static halo using a Monte Carlo approach to simulate an ensemble of clump trajectories. We realized that the heating rate in the core is higher than the cooling rate, which could cause the core to expand. In order to see how the cluster could reach a steady-state configuration one must allow the system to respond dynamically. The current implementation of the model using a 1D hydros code to simulate a cluster in the cosmological context allows us to do just that. We find that convection can regulate the overheating instability and produce a cluster with no cooling flow in steady state. The net effect of the dynamical response is to make the heating more efficient. For example, with 5 per cent of the baryons in $10^8$ M$_\odot$ clumps infalling into a static $3 \times 10^{14}$ M$_\odot$ cluster, our estimates in DB08 indicates a heating to cooling ratio slightly below unity, while here we find it to be above unity, predominantly due to the net expansion of the core. The dynamical evolution of the cluster (Fig. 1) is noticeable especially from $\sim 2$ and on, as the core takes a different thermodynamic trajectory in response to the heating. The BCG mass is smaller, and the core density is lower than in the simulation without heating. It is therefore possible that other heating mechanisms that were tested within a static framework (Fabian 2003; Kim & Narayan 2003; Kim, El-Zant & Kamionkowski 2005; Ciotti & Ostriker 2007; Conroy & Ostriker 2008) would also show different evolution tracks once the gas is allowed to dynamically adjust to the energy input while the halo is growing.

The main missing piece in the proposed model of clumps as the agents for depositing the gravitational energy of infall in the ICM core is the unspecified mechanism and birthplace for the formation of clumps with the desired properties of abundance and mass. While clumps probably do not form in situ at the cores of clusters (Binney et al. 2009), they can form within cosmic filaments (Dekel et al. 2009; Ceverino, Dekel & Bournaud 2010; Fumagalli et al. 2011) and in the edges of haloes. Non-linear perturbations and complicated halo geometries can stimulate the formation of such clumps (Kereš & Hernquist 2009). The analogy between the observed HVCs and the desired clumps for heating clusters is promising and may provide a clue for the origin of these clumps.

The degree of clumpiness needed for effective clump heating, at the level of $\sim 5$ per cent, does not seem to be very demanding. The clumps may be hard to detect outside the halo virial radius as their temperatures are expected to be only slightly lower than the temperature of the surrounding filaments, and therefore their inner densities in pressure equilibrium are expected to be only slightly higher. The clumps become denser and possibly more detectable once they enter the hotter and denser ICM, and especially as they approach the cluster core. We are encouraged by observations of HVC structures with masses of $10^8$–$10^9$ M$_\odot$ around the Perseus BCG.

© 2011 The Authors, MNRAS 415, 2566–2579
Monthly Notices of the Royal Astronomical Society © 2011 RAS
(Fabian et al. 2008) but a more careful comparison needs to be made, addressing the ionization states and the radiative signature of clumps as they are disrupted and heated (Begelman & Fabian 1990; Gnat et al. 2010). One should also work out the spectrum of the X-ray emission from the multiphased gas as it is heated by conduction and radiation. Here we only provide upper and lower limits to the observed temperature and entropy, derived by assuming either that only the hot component is observed, and that full mixing occurs instantaneously as the clumps disintegrate.

The exact details of clump–ICM interactions, and the response of the ICM, cannot be properly addressed in 1D simulations where hydrodynamic instabilities are almost completely suppressed independent of the quenching mechanism. In this first crude study we rely on the fact that the model proposed here naturally deposits the energy over ~1 kpc scales (Fig. 6) in a continuous manner. Nevertheless, a realistic study of whether the gas cooling is sufficiently suppressed would require a proper 3D simulation with clump heating implemented.

The inherent runaway expansion that occurs when heating is faster than cooling is damped by a 1D convection model, which introduces a free mixing length parameter that can only be calibrated by 3D simulations. Our assumption here, that the convection is maximal in the sense that bubbles accelerate until they reach the speed of sound may be overly optimistic. Additionally, weak magnetic fields in the ICM may affect the nature and strength of the convection and could alter the general behaviour of convection to follow temperature inversions rather than entropy inversions. We find that the results are not particularly sensitive to the value of the mixing length parameter as the local perturbations are short scaled and (at least within the framework of the model) entropy inversion is erased for a wide range of mixing length parameters. These effects must be addressed in future work.

In Section 2.1.5 we outlined a proposed implementation of a 3D subgrid model for these clumps, which will allow all these issues to be addressed. A different kind of 2D and 3D simulations, of interactions between single clumps with the ICM gas, have been conducted in the past (Murray & Lin 2004), but not for the specific conditions in clusters, and without some of the crucial physical components such as cooling and conduction.

Our results support the notion that gravitational heating by the instreaming baryons could be a major player in the heating of the cores of massive galaxies and clusters. Whether this mechanism by itself is sufficient for preventing cooling flows or it must work in concert with AGN feedback is yet to be investigated using 3D cluster simulations (e.g. Zinger et al., in preparation).

ACKNOWLEDGMENTS

We thank Ami Glasner for his advice on mixing length theory. We acknowledge fruitful discussions with Peng Oh, Orly Gnat, Jerry Ostriker and Ian Parrish, and thank the referee for helpful comments. This research has been partly supported by ISF grant 6/08, by GIF grant G-1052-104.7/2009, by a DIP grant and by NSF grant AST-1010033. YB was an ITC fellow at CfA, Harvard.

REFERENCES

Balbus S. A., 2000, ApJ, 534, 420
Balbus S. A., 2001, ApJ, 562, 909
Balbus S. A., Reynolds C. S., 2008, ApJ, 681, L65
Begelman M. C., Fabian A. C., 1990, MNRAS, 244, 26r
Bertschinger E., 1985, ApJS, 58, 39
Bertschinger E., Meiksin A., 1986, ApJ, 306, L1
Best P. N., 2007, New Astron. Rev., 51, 168
Binney J., Cowie L. L., 1981, ApJ, 247, 464
Binney J., Nipoti C., Fraternali F., 2009, MNRAS, 397, 1804
Birnboim Y., Dekel A., 2003, MNRAS, 345, 349
Birnboim Y., Loeb A., 2009, J. Cosmol. Astropart. Phys., 6, 8
Birnboim Y., Dekel A., Neistein E., 2007, MNRAS, 380, 339
Birnboim Y., Keshet U., Hernquist L., 2011, MNRAS, 408, 199
Birzan L., Rafferty D. A., McNamara B. R., Wise M. W., Nulsen P. E. J., 2004, ApJ, 607, 800
Blitz L., Spergel D. N., Teuben P. J., Hartmann D., Burton W. B., 1999, ApJ, 514, 818
Bregman J. N., David L. P., 1988, ApJ, 326, 639
Cattaneo A., et al., 2009, Nat, 460, 213
Cavagnolo K. W., Donahue M., Voit G. M., Sun M., 2009, ApJS, 182, 12
Ceverino D., Dekel A., Bournaud F., 2010, MNRAS, 404, 2151
Churazov E., Forman W., Vikhlinin A., Tremaine S., Gerhard O., Jones C., 2008, MNRAS, 388, 1062
Ciotti L., Ostriker J. P., 2007, ApJ, 665, 1038
Conroy C., Ostriker J. P., 2008, ApJ, 681, 151
Conselice C. J., Gallagher J. S., III, Wyse R. F. G., 2001, AJ, 122, 2281
David L. P., Slyz A., Jones C., Forman W., Vrtilek S. D., Arnaud K. A., 1993, ApJ, 412, 479
Dekel A., 1981, A&A, 101, 79
Dekel A., Birnboim Y., 2006, MNRAS, 368, 2
Dekel A., Birnboim Y., 2008, MNRAS, 383, 119(DB08)
Dekel A. et al., 2009, Nat, 457, 451
De Young D. S., O’Neill S. M., Jones T. W., 2008, in Rector T. A., De Young D. S., eds, ASP Conf. Ser. Vol. 386, Extragalactic Jets: Theory and Observation from Radio to Gamma Ray. Astron. Soc. Pac., San Francisco, p. 343
Donahue M., Horner D. J., Cavagnolo K. W., Voit G. M., 2006, ApJ, 643, 739
Edge A. C., Stewart G. C., Fabian A. C., Arnaud K. A., 1990, MNRAS, 245, 559
El-Zant A. A., Kim W.-T., Kambonkowski M., 2004, MNRAS, 354, 169
Fabian A. C., 1994, ARA&A, 32, 277
Fabian A. C., 2003, MNRAS, 344, L27
Fabian A. C., Johnstone R. M., Sanders J. S., Conselice C. J., Crawford C. S., Gallagher J. S. I., Zweibel E., 2008, Nat, 454, 968
Faltenbacher A., Mathews W. G., 2007, MNRAS, 375, 313
Field G. B., 1965, ApJ, 142, 531
Fraternali F., 2008, MNRAS, 386, 935
Fumagalli M., Prochaska J. X., Kasen D., Dekel A., Ceverino D., Primack J. R., 2011, ArXiv e-prints
Giovanelli R. et al., 2007, AJ, 133, 2569
Kent B. R. et al., 2007, ApJ, 665, L15
Kereˇs D., Hernquist L., 2009, ApJ, 700, L1
Khochfar S., Ostriker J. P., 2008, ApJ, 680, 54
Kim W.-T., Narayan R., 2003, ApJ, 596, 889
Kim W.-T., El-Zant A. A., Kambonkowski M., 2005, ApJ, 632, 157
Kunz M. W., Schekochihin A. A., Cowley S. C., Binney J., Sanders J. S., 2011, MNRAS, 410, 2446
Lacey C., Cole S., 1993, MNRAS, 262, 627
Landau L. D., Lifshitz E. M., 1959, Fluid Mechanics. Pergamon Press, New York
Leccardi A., Molendi S., 2008, A&A, 486, 359
Maller A. H., Bullock J. S., 2004, MNRAS, 355, 694
Markevitch M., 1998, ApJ, 504, 27
Murray S. D., Lin D. N. C., 2004, ApJ, 615, 586
Naab T., Johansson P. H., Ostriker J. P., Efstathiou G., 2007, ApJ, 658, 710

© 2011 The Authors, MNRAS 415, 2566–2579
Monthly Notices of the Royal Astronomical Society © 2011 RAS
This paper has been typeset from a LaTeX file prepared by the author.