

UNIFYING THE MICRO AND MACRO PROPERTIES OF AGN FEEDING AND FEEDBACK

MASSIMO GASPARI^{1,*},† & ALEKSANDER SĄDOWSKI^{2,*}

¹Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544-1001 USA; mgaspari@astro.princeton.edu
²MIT Kavli Institute for Astrophysics and Space Research, 77 Massachusetts Ave, Cambridge, MA 02139, USA; asadowsk@mit.edu

Draft version February 21, 2017

ABSTRACT

We unify the feeding and feedback of supermassive black holes with the global properties of galaxies, groups, and clusters, by linking for the first time the physical mechanical efficiency at the horizon and Mpc scale. The macro hot halo is tightly constrained by the absence of overheating and overcooling as probed by X-ray data and hydrodynamic simulations ($\varepsilon_{\text{BH}} \simeq 10^{-3} T_{x,7.4}$). The micro flow is shaped by general relativistic effects tracked by state-of-the-art GR-RMHD simulations ($\varepsilon_{\bullet} \simeq 0.03$). The SMBH properties are tied to the X-ray halo temperature T_x , or related cosmic scaling relation (as L_x). The model is minimally based on first principles, as conservation of energy and mass recycling. The inflow occurs via chaotic cold accretion (CCA), the rain of cold clouds condensing out of the quenched cooling flow and recurrently funneled via inelastic collisions. Within 100s gravitational radii, the accretion energy is transformed into ultrafast 10^4 km s^{-1} outflows (UFOs) ejecting most of the inflowing mass. At larger radii the energy-driven outflow entrains progressively more mass: at roughly kpc scale, the velocities of the hot/warm/cold outflows are a few 10^3 , 1000, 500 km s^{-1} , with median mass rates ~ 10 , 100, several $100 M_{\odot} \text{ yr}^{-1}$, respectively. The unified CCA model is consistent with the observations of nuclear UFOs, and ionized, neutral, and molecular macro outflows. We provide step-by-step implementation for subgrid simulations, (semi)analytic works, or observational interpretations which require self-regulated AGN feedback at coarse scales, avoiding the a-posteriori fine-tuning of efficiencies.

Keywords: black hole accretion – ISM, IGM, ICM – methods: 3D GR-RMHD simulations, analytics

1. INTRODUCTION

Last-decade observations and simulations have shown that supermassive black holes (SMBHs) and cosmic structures are not separate elements of the universe (Heckman & Best 2014 for a review). While cosmic structures are characterized by virial radii¹ r_{100} ($\sim \text{Mpc}$), SMBHs have a characteristic Schwarzschild radius $r_s = 2GM_{\bullet}/c^2$ (10^{-4} pc for $M_{\bullet} = 10^9 M_{\odot}$), implying a difference of 10 dex in length scale. This magnitude of separation might strike as insurmountable, however, black holes would not exist without matter feeding them, and cosmic structures would tend to a quick cold death without feedback from SMBHs (often called active galactic nuclei – AGN – to emphasize such role), thus creating a symbiotic relation.

At the present, no simulation is capable of covering simultaneously the 10 dex dynamic range involving SMBH feeding and feedback (Fig. 1), and to track the evolution from 0.1 yr to 10 Gyr. Recent attempts have been made in the direction of linking the large-scale multiphase gaseous halos of galaxies (ISM), groups (IGM), and clusters (ICM) down to the subpc accretion scale (e.g., Gaspari et al. 2015, 2017 – G15, G17). The dark matter halos heat up the diffuse gas during gravitational collapse, creating stratified hot plasma halos ($\sim 10^7 \text{ K}$) filling cosmic structures, which are detected in X-ray (e.g., An-

derson et al. 2015 and refs. within). Such plasma radiatively cools in the core ($< 0.1 r_{100}$) through a *top-down condensation cascade* to dense warm gas ($\sim 10^4$ - 10^5 K ; optical/IR-UV) and cold gas ($\lesssim 100 \text{ K}$; radio), subsequently raining toward the nuclear region ($< 10^{-3} r_{100}$). Via recurrent collisions, the condensed clouds are rapidly funneled toward the inner stable orbit ($r_{\text{ISCO}} \approx 3 r_s$). Such process is known as *chaotic cold accretion* (Gaspari et al. 2013; §2.1). CCA has been independently probed by several observational works (e.g., Werner et al. 2014; David et al. 2014; Voit et al. 2015; Tremblay et al. 2016 and refs. therein).

General-relativistic, radiative magneto-hydrodynamic simulations (GR-RMHD) provide crucial constraints for the last stage of feeding (e.g., Sądowski et al. 2015, 2016; Sądowski & Gaspari 2017 – SG17; §2.2). Near the ISCO, the final drastic SMBH pull converts a fraction of the gravitational energy into mechanical output, ejecting most of the mass via *ultra-fast outflows* (UFOs). Such outflows re-heat the core, while entraining the ambient gas, in a self-regulated AGN feedback loop (Fig. 1). In the paper companion to this work (SG17), we present and discuss in-depth the GR-RMHD simulations results, including the mechanical and radiative efficiencies.

In §3, we will quantitatively link the macro and micro properties of cosmic structures and SMBHs by using first principles, as mass and energy conservation, and by preserving minimal assumptions based on last-decade observations. The final equations provide the mass outflow rates and velocities at different scales (and for different phases). In §4, we compare the predictions with recent ionized, warm, and molecular outflow samples, and discuss the limitations. In §5, we discuss how to apply

* Einstein Fellow.

† Spitzer Fellow.

¹ The radius r_{δ} encloses δ times the critical overdensity $\rho_c(z) = 3H^2(z)/(8\pi G)$ (H is the Hubble parameter; $H_0 \simeq 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$) giving an enclosed mass $M_{\delta} = (4\pi/3)\delta\rho_c(z)r_{\delta}^3$; $\delta \simeq 100$ for the virial radius and 500 for observational constraints.

our model to other studies, as subgrid simulations, semi-analytic (SAM) studies, or observational interpretations. In §6, we carefully discuss the limitations of the model and additional important features (as the duty cycle and the $M_\bullet - \sigma_*$ relation). In §7, we summarize the main points and conclude with future prospects.

2. LARGE AND SMALL SCALES EFFICIENCIES

We highlight here 3 key regions which are central to our study (see Fig. 1 for a full diagram).

(i) The region closest to the SMBH *horizon*, $r \lesssim 5 r_S$ (a few ISCO radii), where gas is rushing toward the BH and there is no outflow. This region is fully resolved by the horizon scale GR-RMHD simulations. We denote properties in this region by a black dot, e.g., \dot{M}_\bullet .

(ii) The ultra-fast outflow *launching* region, $r \lesssim 100 r_S$, within which the binding energy of the infalling gas is converted into mechanical outflow, not interacting yet with the ambient gas. We denote this by \dot{M}_{out} .

(iii) The *macro* region, $r < r_{\text{core}} \approx 10^9 r_S \approx 0.1 r_{\text{vir}}$, within which the nuclear outflow is entrained (denoted by \dot{M}_{OUT}), slowed down, and eventually thermalized (via bubbles, shocks, and turbulent mixing). The CCA rain develops in such core, with major collisions increasing within the kpc scale (10-100 Bondi radii²).

2.1. Macro efficiency: chaotic cold accretion [CCA]

We introduce now the two key property of the feeding and feedback, i.e., the mechanical efficiency which has dimension of a power divided by the rest mass energy rate, $\varepsilon \equiv P/(\dot{M}c^2)$.

The best consistent way to solve the cooling flow problem appears to be mechanical AGN feedback self-regulated via CCA (§1). Solving the cooling flow problem means to avoid at the same time overcooling and overheating, preserving the inner structure of hot halos for ~ 10 Gyr, as tightly constrained over the last decade by *Chandra* and *XMM-Newton* data (e.g., [McNamara & Nulsen 2012](#); [McDonald et al. 2016](#)). Such hot halos are continuously perturbed by subsonic turbulence (e.g., [Khatri & Gaspari 2016](#)). In turbulent regions where the cooling time drops below the local dynamical time, nonlinear multiphase condensation develops (Fig. 1, bottom). Such cold clouds and warm filaments collide in chaotic, inelastic way while raining on the SMBH (G15, G17; see also [Pizzolato & Soker 2010](#)), boosting the accretion rate with rapid intermittency. Massive sub-relativistic outflows are then triggered with kinetic power P_{OUT} proportional to the large-scale inflow rate, preventing a run-away pure cooling flow (§3).

Due to self-regulation, the large-scale mechanical efficiency can be estimated by comparing the AGN energy output with the radiative energy losses, $P_{\text{OUT}} \simeq L_x$, yielding (§3.1 for the derivation)

$$\varepsilon_{\text{BH}} \simeq 10^{-3} T_{x,7.4} \propto \frac{c_{s,x}^2}{c^2}, \quad (1)$$

where $c_{s,x}$ is the hot halo adiabatic sound speed and c is the speed of light (the scaling shares analogy to a Mach

² The Bondi radius, $r_B = GM_\bullet/c_{s,x}^2 \simeq (7.5 \text{ pc}) M_{\bullet,9} T_{x,7.4}^{-1} \approx 10^5 r_S$, is not strictly relevant for CCA but provides a known reference intermediate (pc) scale between the macro and micro region.

number squared). From less massive, lower-temperature, compact galaxies to more massive, hotter, and larger galaxy clusters ($T_x \approx 0.5 - 10$ keV), the mechanical efficiency covers a range $\varepsilon_{\text{BH}} \simeq 2 \times 10^{-4} - 4 \times 10^{-3}$. The macro efficiency is a function of hot halo temperature ($\propto T_x$), thus total mass, decreasing for smaller halos since the cooling rate is a function $\propto L_x/T_x$ (as seen later in Eq. 6). Smaller, less bound halos experience a stronger relative condensation due to the lower specific internal energy, and necessitate of less sinked material – with slightly more evacuation – in order to avoid over-heating. Such quasi thermal equilibrium constraint on X-ray halos filling cosmic systems is key to set the macro efficiency.

This picture has been corroborated by self-regulated AGN simulations of CCA and massive outflows tested in clusters, groups, and isolated galaxies (e.g., [Gaspari et al. 2011a,b](#), [2012a,b](#); [Prasad et al. 2015](#); [Yang & Reynolds 2016](#)), which independently retrieve the same range of feedback efficiencies described above in varying systems. The few available observational estimates, albeit limited by several extrapolations, are also consistent with a mechanical efficiency of the order of $\varepsilon_{\text{BH}} \sim 10^{-3}$ ([Merloni & Heinz 2008](#)).

2.2. Horizon efficiency: GR-RMHD

Gas approaching the SMBH liberates its gravitational energy. A test particle falling straight on the BH would convert the liberated amount into kinetic energy of radial motion and, finally, take it with it below the horizon. From the point of view of the observer at infinity, no energy has been extracted. Accretion flows act in a more complex way. The liberated gravitational energy goes mostly into kinetic motion. The turbulent nature of the flow induces this energy to dissipate and heat up the gas. At the same time outflows can be generated often via the magneto-centrifugal mechanism. Only for idealized models, like advection dominated accretion flows (e.g., [Narayan & Yi 1995](#)), all the dissipated heat is advected with the flow onto the BH. In a more general case, energy is extracted from the system, and gas infalling from large radii and marginally bound, crosses the BH horizon with negative energy.

The amount of the extracted energy, i.e., the efficiency of a given accretion flow, depends solely on the energetics of the magnetized gas crossing the BH horizon. E.g., if on average gas with energy $e = -0.01 \rho c^2$ falls into the BH, then the luminosity of such a system, as seen from infinity, is $L = 0.01 \dot{M}_\bullet c^2$. The properties of the accretion flow in the innermost region must be determined by numerical means, since the flow is highly nonlinear, strongly magnetized, and turbulent. In the companion paper, SG17, general relativistic radiative simulations of magnetized gas falling on the SMBH are carried out, testing over 5 orders of magnitude in accretion rates. SG17 show that for a non-rotating BH and standard non-saturated configuration of the magnetic field, thick accretion flows (as expected in the maintenance mode of AGN feedback) have a fairly stable extraction of the rest mass energy accreted through the horizon,

$$\varepsilon_\bullet \simeq 0.03 \pm 0.01, \quad (2)$$

Such mechanical efficiency will be the reference horizon

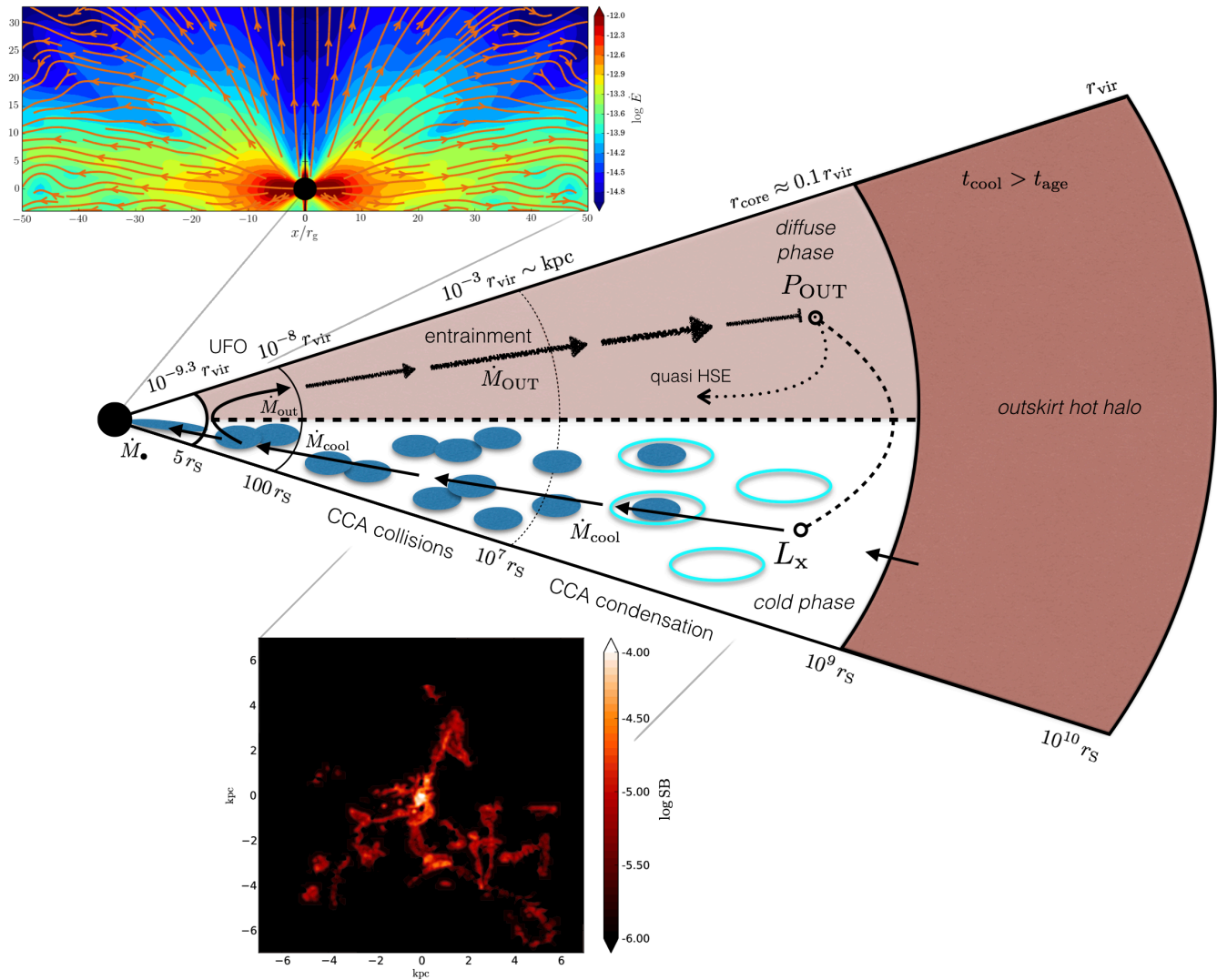


Figure 1. *Middle* – Diagram of the multiphase accretion inflow and outflow covering the entire range of scales, from the inner SMBH horizon to the virial radius of the galaxy, group, or cluster. The self-regulated AGN feedback loop works as follows. The turbulent gaseous halo condenses in localized, large-scale high density peaks (cyan), leading to the drop out of cold clouds and warm filaments (blue). The clouds rain down and recurrently collide in chaotic and inelastic way (CCA), canceling angular momentum and flowing toward the SMBH. The mass inflow rate originates from the *quenched* X-ray cooling rate within the core region. Within $\sim 100 r_S$, the gravitational accretion process releases ultrafast outflows (UFOs), while only a small gas fraction is sunk through the horizon (this is balanced by a net inflow from the outskirts). The outflows slow down at larger radii, entraining the gas of the background profile. The energy is thermalized in the core, balancing the X-ray luminosity. The CCA rain is thus stopped, and so are the outflows, allowing the global halo to restore the quasi HSE profile. As cooling resumes without a source of heating, another cycle of CCA rain and collisions, mass ejection and entrainment, and restoration is triggered, consistently with X-ray data. The system conserves total energy and mass in a gentle recycling multiphase flow. *Top* – GR-RMHD simulation of the micro flow (§2.2), showing the magnitude and streamlines of the total energy flux (from Sądowski et al. 2016; code units) which is dominated by the kinetic component with $\epsilon_* \simeq 0.03$ (see SG17 for more details on the mechanical efficiency). *Bottom* – Multiphase hydro simulation of the macro flow tracking the CCA evolution (from Gaspari et al. 2017; §2.1). The map shows the surface brightness ($\text{erg s}^{-1} \text{cm}^{-2} \text{sr}^{-1}$) of the filamentary warm phase condensed out of the turbulent hot halo and chaotically colliding.

efficiency for our model. We note that chaotic accretion (our macro scale model) will naturally lead to an average null spin configuration (e.g., King & Pringle 2006). An important result from SG17 is that such value is essentially independent of the ion-electron temperature ratio, i.e., the strength of the gas cooling does not affect the mechanical efficiency value at the micro scale.

This energy outflow accelerates within the inner region ($\sim 100 r_S$) and is ejected in a quasi-spherical way (Fig. 1, top) in the form of an ultra-fast kinetic outflow of gas. The outflow is both thermally (equatorial) and magnetically driven (polar region; see also the simulations by Moller & Sądowski 2015). At larger distances,

the outflow interacts with the ambient medium, entraining gas via shocks and mixing instabilities, finally dissipating its energy within the core region, $r_c \approx r_{500}/5 = (148 \text{ kpc}) T_{x,7.4}^{1/2}$ (App. A). On top of this energy flux there might be a very thin, relativistic jet forming whenever the SMBH quickly spins and the magnetic field threads the horizon. Such a jet may be substantially energetic and could lead to larger efficiencies (e.g., Tchekhovskoy et al. 2011). However, relativistic jets are in most cases very collimated and less likely to interact with the host. For such reason and for the null spin expected from chaotic accretion, we consider here only the wide, mas-

sive sub-relativistic outflows as dominant component of the kinetic feedback. Albeit not driving the total ram pressure, we note the jet and radio emission can still be correlated with the presence of massive AGN outflows, thus tracing some of the major AGN bubbles (§6).

The emergence of ultra-fast outflows and the connection with large-scale warm absorbers has been corroborated by other analytic studies. Fukumura et al. (2010, 2014) show that magnetic torques acting on the inner rotating gas can efficiently drive an outflow through the magneto-centrifugal mechanism. The MHD wind is stratified, having slower velocity at progressively larger launching radii, akin to an entrained outflow. In the radiatively efficient, Eddington regime, the spherical model by King & Pounds (2014) suggests that radiation pressure is able to drive UFOs; the expanding, swept-up shell is decelerated by the background medium, again corroborating the key role of the entrainment action in unifying AGN outflows over a large range of scales.

3. LINKING THE MACRO AND MICRO SCALES

The two complementary simulations discussed above allow us to link the large-scale to small-scale properties of the feeding and feedback mechanism in a simple, coherent model. Fig. 1 illustrates the main features and characteristic scales of the model.

The large-scale outflow power can be modeled as

$$P_{\text{OUT}} = \varepsilon_{\text{BH}} \dot{M}_{\text{cool}} c^2, \quad (3)$$

where \dot{M}_{cool} is the quenched cooling flow rate and ε_{BH} is the macro-scale mechanical efficiency (§2.1). The gaseous halo is losing internal energy via radiative emission (mainly via Bremsstrahlung), while the AGN feeds heating back, on average balancing the pure cooling flow. Such halos perturbed by subsonic turbulence develop local multiphase condensation within the core, as long as turbulent Taylor number $\text{Ta}_t \equiv \sigma_v/v_{\text{rot}} \lesssim 1$ (G15). As cold clouds and filaments rain down, they experience recurrent chaotic, fractal collisions, which cancel angular momentum at progressively smaller radii, in particular as they collapse within $r < 1$ kpc. The inflow rate can be thus considered independent of radius. In other words, during CCA rain, the cold gas condensed in the core is quickly funneled to the ISCO region with no long-term accumulation. G17 simulations showed that the CCA inflow rate is proportional to the effective viscosity of the cloud collisions, $\nu_c \approx \lambda_c \sigma_v$. The collisional mean free path λ_c and the ensemble velocity dispersion σ_v are directly inherited from the large-scale turbulence (for a massive galaxy, $\lambda_c \approx 100$ pc, $\sigma_v \approx 150$ km s⁻¹) – a 3D chaotic process not tied to a radial dependence.

The inner, tiny SMBH is the actual source of energy injection with power

$$P_{\text{out}} = \varepsilon_{\bullet} \dot{M}_{\bullet} c^2, \quad (4)$$

where ε_{\bullet} is the horizon efficiency (§2.2) and \dot{M}_{\bullet} is the inflow rate through the black hole horizon. The major difference between the macro and horizon efficiency implies that the sinked mass rate is the net inflow rate surviving the ultra-fast outflow generated near the ISCO scale, before falling into the unescaping BH horizon.

3.1. Inflow properties

The large-scale inflow rate is the *quenched* cooling flow rate. The maximal *pure* cooling flow (CF) rate can be calculated from the enthalpy variation of the hot gaseous halo via the first law of thermodynamics (e.g., Gaspari 2015) in isobaric mode, yielding

$$L_x = \frac{\gamma}{\gamma - 1} \frac{k_b T_x}{\mu m_p} \dot{M}_{\text{CF}} = \frac{c_{s,x}^2}{\gamma - 1} \dot{M}_{\text{CF}}, \quad (5)$$

where T_x and L_x denote the core X-ray temperature and luminosity of the hot halo (App. A), $\gamma = 5/3$ is the adiabatic index, $\mu \simeq 0.62$ is the average atomic weight for a fully ionized plasma with $\simeq 25\%$ He in mass, and k_b and m_p are the usual Boltzmann constant and proton mass, respectively. The last equality converts temperature into adiabatic sound speed, $c_s = (\gamma k_b T_x / \mu m_p)^{1/2} \simeq 1.5 \times 10^4 T_x^{1/2}$. From galaxies to clusters ($T_x \approx 0.5 - 10$ keV), $c_{s,x} = 361 - 1615$ km s⁻¹.

AGN feedback preserves the hot halos in quasi thermal equilibrium throughout the 10 Gyr evolution³. The warm filaments drop out of the hot halo just below the soft X-ray regime (G17) as the cooling curve drastically increases due to line cooling. Thereby the actual mass flux arising out of the condensation process is linked to the suppressed soft X-ray luminosity. X-ray spectroscopical observations (e.g., Peterson et al. 2003; Kaastra et al. 2004) show that the soft X-ray emission is on average suppressed by 2 dex compared with the pure isobaric CF tied to the core L_x (cf. Gaspari 2015 for a review of observational works and analysis of the soft X-ray spectrum quenching). For such reasons, the effective quenched cooling rate is

$$\dot{M}_{\text{cool}} \simeq 10^{-2} \dot{M}_{\text{CF}} \simeq 6.7 \times 10^{-3} \frac{L_x}{c_{s,x}^2}. \quad (6)$$

We note such quenched, CCA rate is typically 100× the Bondi rate (Gaspari et al. 2013), the latter being insufficient to properly boost the AGN heating (see also Soker et al. 2009; McNamara & Nulsen 2012). Since hot halos are formed via the gravitational collapse of the cosmic structures, the temperature and luminosity are interchangeable via scaling relations (Sun 2012), such as $L_x \simeq 6 \times 10^{43} (T_x/2.2 \text{ keV})^3 \text{ erg s}^{-1}$ (including the minor corrections due to the core radius instead of R_{500} ; see App. A). We can thus rewrite Eq. 6 as

$$\dot{M}_{\text{cool}} \simeq (1.1 \text{ M}_{\odot} \text{ yr}^{-1}) T_{x,7.4}^2 = (1.1 \text{ M}_{\odot} \text{ yr}^{-1}) L_{x,43.8}^{2/3}, \quad (7)$$

where the core L_x and T_x are in unit of $6 \times 10^{43} \text{ erg s}^{-1}$ and $2.6 \times 10^7 \text{ K}$ (2.2 keV), respectively. From compact galaxies to massive clusters ($T_x \approx 0.5 - 10$ keV), the inflow rate covers $\dot{M}_{\text{cool}} \simeq 0.06 - 23 \text{ M}_{\odot} \text{ yr}^{-1}$. Interestingly, all the below scalings can be also expressed in terms of total mass or virial radius, e.g., $\dot{M}_{\text{cool}} \propto r_{\text{vir}}$ (App. A). It is important to note that if the core cooling time is $t_{\text{cool}} \gtrsim t_{\text{H}}/2$, then the system is in a non-cool-core condition and no condensation rain, feeding, and feedback shall be applied (regardless of scaling relations), until the core cools down, igniting the self-regulated loop (see §6).

³ McDonald et al. (2017) show that cool cores are observed even up to $z \approx 1.5$ with properties identical to local ones.

The energy conservation requirement,

$$P_{\text{out}} = P_{\text{OUT}}, \quad (8)$$

implies that the horizon inflow rate is related to the cooling rate as follows:

$$\dot{M}_{\bullet} = \frac{\varepsilon_{\text{BH}}}{\varepsilon_{\bullet}} \dot{M}_{\text{cool}}, \quad (9)$$

where the horizon mechanical efficiency is directly provided by the GR-RMHD simulations (§2.2), $\varepsilon_{\bullet} = 0.03$. From the results and observations discussed in §2.1, hot halos must avoid at the same time overheating and overcooling, i.e., the energy lost via radiative emission in the core must be replaced by the SMBH feedback power,

$$P_{\text{OUT}} \simeq L_{\text{x}}. \quad (10)$$

Thereby $\varepsilon_{\text{BH}} = L_{\text{x}}/(\dot{M}_{\text{cool}}c^2)$ and by using Eq. 6, the macro efficiency reduces to

$$\varepsilon_{\text{BH}} = \frac{10^2}{\gamma - 1} \frac{c_{\text{s}}^2}{c^2} \simeq 10^{-3} T_{\text{x},7.4} = 10^{-3} L_{\text{x},43.8}^{1/3} \quad (11)$$

Notice that the efficiency only mildly varies with the main variable, the X-ray luminosity. We can now use both efficiencies to retrieve the horizon inflow rate relative to the macro value via Eq. 9 as

$$\dot{M}_{\bullet} \simeq (0.03 \dot{M}_{\text{cool}}) T_{\text{x},7.4} = (0.03 \dot{M}_{\text{cool}}) L_{\text{x},43.8}^{1/3}, \quad (12)$$

i.e., only a few percent of the quenched cooling flow rate is actually sinked through the SMBH horizon. Substituting \dot{M}_{cool} in Eq. 12, we consistently retrieve the accretion rate directly proportional to the X-ray luminosity,

$$\begin{aligned} \dot{M}_{\bullet} &= \frac{L_{\text{x}}}{\varepsilon_{\bullet} c^2} \simeq (0.04 M_{\odot} \text{yr}^{-1}) L_{\text{x},43.8} \\ &= (0.04 M_{\odot} \text{yr}^{-1}) T_{\text{x},7.4}^3. \end{aligned} \quad (13)$$

For SMBHs in the local universe, such accretion rates are typically sub-Eddington, as expected for the maintenance, mechanically dominated mode of AGN feedback. As shown by Russell et al. (2013) and corroborated by SG17, the radiative efficiency and thus power due to radiation is several dex lower than the mechanical input, and it can be neglected in terms of driver of the dynamics (albeit radiation is clearly relevant to detect and trace AGN; §6). Eq. 12-13 imply that SMBHs in lower mass halos have typically a lower *absolute* accretion rate. Moreover, a *relatively* smaller fraction of gas reaches the horizon as AGN feedback is more effective in halos with lower binding energy, which are tied to both lower M_{500} and lower black holes masses (§6).

3.2. Outflow properties

Having assessed the inflow properties, we are now in a position to retrieve the structure of the outflows, again via minimal first principles. The power in terms of characteristic mass outflow rates⁴ and velocities at the launching and macro scale is

$$P_{\text{out}} = \frac{1}{2} \dot{M}_{\text{out}} v_{\text{out}}^2, \quad (14)$$

⁴ The term due to $v\dot{v}$ is subdominant and can be neglected.

$$P_{\text{OUT}} = \frac{1}{2} \dot{M}_{\text{OUT}} v_{\text{OUT}}^2, \quad (15)$$

respectively. As shown in Eq. 12, only a few percent of the total inflow is actually sinked through the SMBH horizon; most of the mass is returned as ultra-fast outflows launched within $\sim 100 r_{\text{S}}$, such as

$$\dot{M}_{\text{out}} = \dot{M}_{\text{cool}} - \dot{M}_{\bullet} = \left(1 - \frac{\varepsilon_{\text{BH}}}{\varepsilon_{\bullet}}\right) \dot{M}_{\text{cool}} \approx \dot{M}_{\text{cool}}, \quad (16)$$

which leads to the inner outflow velocity via Eq. 14

$$\begin{aligned} v_{\text{out}} &= \sqrt{\frac{2\varepsilon_{\bullet} \dot{M}_{\bullet} c^2}{\dot{M}_{\text{out}}}} = \sqrt{\frac{2\varepsilon_{\text{BH}}}{1 - \varepsilon_{\text{BH}}/\varepsilon_{\bullet}}} c \simeq \sqrt{2\varepsilon_{\text{BH}}} c \\ &\simeq (1.4 \times 10^4 \text{ km s}^{-1}) T_{\text{x},7.4}^{1/2} = (1.4 \times 10^4 \text{ km s}^{-1}) L_{\text{x},43.8}^{1/6}. \end{aligned} \quad (18)$$

We note v_{out} in Eq. 17 can be tied to a momentum $p_{\text{out}} = \dot{M}_{\text{out}} v_{\text{out}}$, which satisfies $(1/2)\dot{M}_{\text{out}}v_{\text{out}}^2 = p_{\text{out}}^2/(2\dot{M}_{\text{out}})$.

Together with the above outflow rates, these are the typical velocities of ultra-fast outflows (UFOs) observed as blue-shifted absorption lines tracing the inner launching region near the SMBH gravitational radius (Tombsi et al. 2012, 2013; Fukumura et al. 2015; more discussions and comparisons in §4). We note the outflow velocity is only weakly dependent on the halo temperature/luminosity, varying at best by a factor of 2.5. We thus expect 10^4 km s^{-1} to be a fairly general attribute⁵ of inner launching outflows (Crenshaw et al. 2003; Tombsi 2016 for reviews).

As the inner ultra-fast outflow propagates outward ($r \gg 100 r_{\text{S}}$), it will entrain the *background* gas (embedding the low volume-filling CCA rain⁶) along its way such as

$$\dot{M}_{\text{OUT}} = \eta \dot{M}_{\text{out}}, \quad (19)$$

where $\eta > 1$ is the entrainment factor. We note at kpc scale the mechanical outflow has not yet thermalized, conserving most of the kinetic energy, as we see the formation of X-ray cavities and hot spots at larger distances. At a given radius, the entrained mass outflow rate can be retrieved via the mass flux equation

$$\begin{aligned} \dot{M}_{\text{OUT}} &= \Omega r^2 \rho(r) v_{\text{OUT}}(r) = \Omega r^{2-\alpha} \rho_0 r_0^{\alpha} v_{\text{OUT}}(r) \\ &\simeq \Omega \rho_0 r_0 r v_{\text{OUT}}, \end{aligned} \quad (20)$$

where the inner gas density profile is typically a power-law $\rho = \rho_0 (r/r_0)^{-\alpha}$ and $\Omega \leq 4\pi$ is the covering angle of the bipolar outflow. As shown in G17 and observational refs. within, the typical nuclear profiles for all the phases follow a slope $\alpha \simeq 1$ (with ≈ 0.25 scatter), hence the last step in Eq. 20. By using Eq. 15 and 19, the entrained outflow velocity can be written as

$$v_{\text{OUT}} = \sqrt{\frac{2P_{\text{OUT}}}{\dot{M}_{\text{OUT}}}} = \eta^{-1/2} v_{\text{out}}, \quad (21)$$

⁵ This is also similar to the characteristic nuclear (100-200 r_{S}) escape velocity, i.e., as the driven outflow overcomes gravity.

⁶ Through the feedback cycle, the underlying halo gently expands during entrainment, and contracts after dissipation, restoring quasi hydrostatic equilibrium (HSE). X-ray observations indeed show that density profiles in cool-core systems vary only by a small amount, even after strong outbursts (e.g., McNamara et al. 2016).

which inserted in Eq. 20 yields an entrainment factor

$$\eta = \left(\Omega \rho_0 r_0 r \frac{v_{\text{out}}}{\dot{M}_{\text{out}}} \right)^{2/3} \propto \frac{r^{2/3}}{T_x}. \quad (22)$$

This implies that, while the macro velocities at a given radius are unchanged over different systems ($v_{\text{OUT}} \propto T_x^{1/2}/T_x^{1/2}$), and are thus more robust probes, the macro outflow rate linearly increases for more massive systems $\dot{M}_{\text{OUT}} \approx \eta \dot{M}_{\text{cool}} \propto T_x$. Note the mass outflow rate has much stronger relative variations than velocities ($\propto \eta^{-1/2}$), corroborating Eq. 14-15.

Depending on the current thermodynamical background state of the system, the outflows can entrain different phases, including the hot plasma, the warm neutral/ionized gas, and the molecular gas. We use the results of the CCA simulations (G17) to retrieve the multiphase environment and profiles of the 3 phases, taking as reference macro scale $r_0 = 1$ kpc. A typical plasma density $\rho_{0,\text{hot}} \simeq 10^{-25} \text{ g cm}^{-3}$ at 1 kpc leads to an entrainment factor ($\Omega \simeq 4\pi$)

$$\eta_{\text{hot}} \simeq 40 T_{x,7.4}^{-1} r_{1 \text{ kpc}}^{2/3}. \quad (23)$$

This implies median entrained mass outflow rates and velocities of $10s M_\odot \text{ yr}^{-1}$ and a few 10^3 km s^{-1} , which are typical properties of observed macro ionized outflows (e.g., Nesvadba et al. 2010; Tombesi et al. 2013). If the halo is mainly filled with cooler gas, such as at high redshift, the entrainment can also proceed mainly via the warm ($\rho_{0,\text{warm}} \simeq 10^{-24} \text{ g cm}^{-3}$) and cold phase ($\rho_{0,\text{cold}} \simeq 10^{-23} \text{ g cm}^{-3}$)⁷, thus leading to more entrained outflows with

$$\eta_{\text{warm}} \simeq 183 T_{x,7.4}^{-1} r_{1 \text{ kpc}}^{2/3}, \quad (24)$$

$$\eta_{\text{cold}} \simeq 850 T_{x,7.4}^{-1} r_{1 \text{ kpc}}^{2/3}. \quad (25)$$

Mass outflow rates with 10^2 and several $10^2 M_\odot \text{ yr}^{-1}$ tied to velocities 1000 and 500 km s^{-1} at the kpc scale are characteristic properties found throughout observations of neutral (e.g., Morganti et al. 2005, 2007; Teng et al. 2013; Morganti 2015) and molecular AGN outflows (Sturm et al. 2011; Ciccone et al. 2014; Russell et al. 2014; Combes 2015; Feruglio et al. 2015; Morganti et al. 2015; Tombesi et al. 2015), respectively (more detailed comparisons in §4).

At large radii, the outflow is halted by the external pressure, inflating a bubble and thermalizing its kinetic energy mainly via turbulent mixing (e.g., Gaspari et al. 2012a; Soker 2016; Yang & Reynolds 2016). Such radius crudely corresponds to the region where the outflow ram pressure becomes equal to the hot halo pressure. Since outflow ram pressure is equal for all the phases, we can estimate the thermalization radius as $v_{\text{OUT,hot}}^2 \sim c_s^2/\gamma$, yielding via Eq. 21-23

$$r_{\text{th}} \sim (55 \text{ kpc}) T_{x,7.4}^{3/2} = (55 \text{ kpc}) L_{x,43.8}^{1/2}. \quad (26)$$

⁷ Here we assume that the characteristic phase densities retrieved in G17 apply over the whole inner region as a background; this is more typical at high redshift, as cold flows can penetrate deep within the growing proto-galaxy.

Above such thermalization radius, any model should simply inject thermal energy rate balancing the core L_x . Below such radius (as resolved by most of the current MHD and cosmological simulations), any model should inject massive outflows with the above relations. Such radius roughly approaches the core radius, which is where the feedback loop is active.

In principle the momentum equation, $\dot{M}_{\text{OUT}} v_{\text{OUT}} = \dot{M}_{\text{out}} v_{\text{out}}$, might be adopted instead of Eq. 14-15, if the outflow would immediately radiate away most of its energy. However, besides losses being likely subdominant (see Faucher-Giguère & Quataert 2012), the deceleration would result to be dramatic, $v_{\text{OUT}} = v_{\text{out}}/\eta$ (with $\eta \propto r^{1/2} T_x^{-3/4}$ reduced by a few), which would make the outflow aborted at the macro scale, inconsistently with data. Adopting the same procedure as above, the hot, warm, and molecular outflow would merely preserve 870, 280, and 90 km s^{-1} at 1 kpc scale, respectively. A related crucial point to reject purely momentum-driven outflows is that self-regulation would be broken, since the macro feedback energy could not balance the core L_x , leading to a global massive pure CF.

4. COMPARISON WITH OBSERVATIONS

The proposed CCA GR-RMHD unification predicts nuclear ultra-fast outflows of the order of 10^4 km s^{-1} and a progressively slower propagation of the outflow at larger radii, which are consistent with recent AGN data.

In a sample of 35 AGN, Tombesi et al. (2013) unify the velocities of UFOs and the slower warm absorbers as a function of radial distance (see also Tombesi et al. 2014 for analogous radio galaxy sample). Velocity is the most robust indicator (e.g., compared to mass outflow rates) since directly observed through blue-shifted absorption lines in AGN X-ray spectra. Fig. 2 shows the comparison of our model prediction (blue; §3) and the fit to the unified X-ray data. The bands denote 0.5 dex scatter, which is the typical model variation (mainly due to inner density and bipolar angle) and the range in the observed data points. The prediction of the CCA GR-RMHD model well reproduces the observed values. If the outflow would be purely driven by momentum (green line) and not energy, it would be aborted within the Bondi radius, remaining clearly below data. In other words, entrainment must occur in a gentle way, such as $v_{\text{OUT}} \propto \eta^{-1/2} \propto r^{-1/3}$. In the nuclear region, the outflow tends to be slightly lower than the data, albeit within typical uncertainties. The slope of the data, -0.40, is slightly steeper than the -0.33 model. The two matches exactly if the density profile has slightly shallower $\alpha = 0.8$ (instead of 1); we did not attempt to fine-tune it, since within uncertainties of the simulated radial profiles and not granting further insight.

The mass outflow rates have very large observational uncertainties (due to the unknown geometry and projection effects) and theoretical scatter (due to the T_x dependence, unlike the macro velocity). In the above sample, UFOs typically show $\dot{M}_{\text{out}} \approx 0.3 M_\odot \text{ yr}^{-1}$, while the warm absorbers have 1.5-2 dex larger magnitude, which can be explained via the entrainment action ($T_x \approx 0.6 \text{ keV}$). We are here not attempting to fit values of single objects; nevertheless, several X-ray studies detect nuclear 10^4 km s^{-1} UFOs and ionized outflows with 10^3 km s^{-1} at

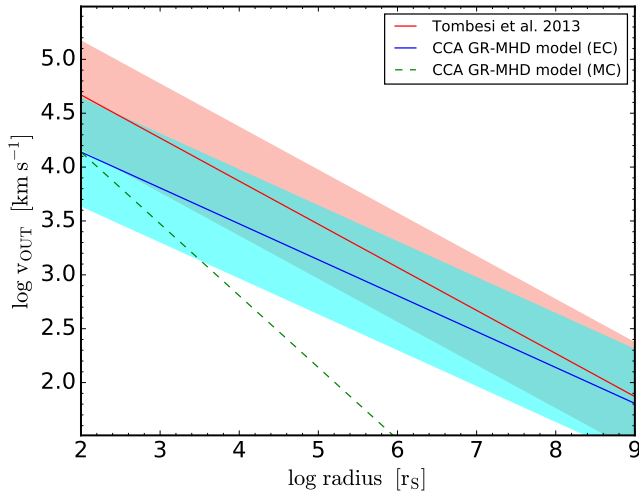


Figure 2. Outflow velocity as a function of radial distance (normalized to the Schwarzschild radius) for the unified X-ray UFO plus warm absorber data (red; Tombesi et al. 2013) and the prediction of our energy-conserving CCA GR-RMHD model (blue; §3). The dashed green line shows the (inconsistent) purely momentum-driven outflow. The region within $\sim 100 r_s$ is the UFO generation region, where most of the inflow mass is ejected. At larger radii, the UFO entrains progressively more mass, slowing down. The adopted profile slope of the warm gas background is $\alpha = 1$. The proposed model, based on linking the horizon/GR-RMHD and macro/CCA efficiencies, well reproduces the data within scatter.

intermediate scale down to several 100 km s^{-1} at large radii⁸ (see the review by Tombesi 2016 and refs. within). Follow-up observational investigations are required to better unify the radial properties of ionized outflows over a large homogeneous sample, in particular adding more low-luminosity AGN and central galaxies.

Depending on the dominant nuclear phase, the AGN ejecta can also develop into a neutral and molecular outflow. This is more common in QSOs and ULIRGs with abundant cold/warm mass with large volume filling in the core. Morganti et al. (2005, 2007) have shown the incidence of HI outflows in several AGN, particularly radio-loud sources, via (21-cm) radio telescopes as WRST. The location of the HI outflows is 0.5 - 1.5 kpc with average velocities 1000 km s^{-1} . Teng et al. (2013) present a sample of 27 kpc-scale HI outflows detected with GBT: the average sample velocity is $v_{\text{OUT,neu}} \simeq 885 \text{ km s}^{-1}$. In both samples, the mass outflow rates are uncertain (due to the dynamical time estimate), of the order of $100 M_{\odot} \text{ yr}^{-1}$. The above values are consistent with our median prediction of neutral outflows (Eq. 21-24) with a typical $v_{\text{OUT,neu}} \simeq 1035 \text{ km s}^{-1}$ and $\dot{M}_{\text{OUT,neu}} \simeq 92 M_{\odot} \text{ yr}^{-1}$ at kpc scale ($T_x \approx 1 \text{ keV}$).

In the last several years and with the advent of high-resolution radio interferometers, neutral outflows have been complemented with samples of massive molecular AGN outflows. Cicone et al. (2014) present a sample of 19 molecular AGN outflows detected with IRAM (by using CO[1-0] emission closely tracing H_2 gas) at the kpc scale. Averaging the peak velocity and mass outflow rates over the sample yields a velocity $v_{\text{OUT,mol}} \simeq 573 \text{ km s}^{-1}$ and mass rate $\dot{M}_{\text{OUT}} \simeq 428 M_{\odot} \text{ yr}^{-1}$ with factor of 2 uncertainty. The sample of 6 molecular out-

flows in Sturm et al. (2011) show very similar mean properties. From Eq. 25, the average molecular velocity and mass outflow rate at kpc scale is expected to be 480 km s^{-1} and $\dot{M}_{\text{OUT}} \simeq 425 M_{\odot} \text{ yr}^{-1}$ ($T_x \approx 1 \text{ keV}$), in agreement with the data. Other works focus on single objects, finding very similar properties at kpc scale as predicted by our model; e.g., Phoenix/A1664 BCG cores display $v_{\text{OUT,mol}} \simeq 550/590 \text{ km s}^{-1}$ and crude outflow rates $> 250 M_{\odot} \text{ yr}^{-1}$ (Russell et al. 2014, 2016). A well studied multiphase outflow in both the hot and cold phase is Mrk 231 (Feruglio et al. 2010, 2015). IRAM data indicates a kpc-scale molecular outflow with $v_{\text{OUT,mol}} \simeq 750 \text{ km s}^{-1}$ and $\dot{M}_{\text{OUT}} \simeq 700 M_{\odot} \text{ yr}^{-1}$ (Feruglio et al. 2010); in the same system, Chandra and NuSTAR show the presence of a nuclear UFO with $v_{\text{out,hot}} \simeq 2 \times 10^4 \text{ km s}^{-1}$ and $\dot{M}_{\text{OUT}} \simeq 1 M_{\odot} \text{ yr}^{-1}$. Both values are in excellent agreement with our entrainment multiphase model. Notably, the same authors remark that energy is conserved during the entrainment process, $P_{\text{out}} \approx P_{\text{OUT}}$, consistently with our Eq. 8. Tombesi et al. (2015) present another similar multiphase outflow in IRAS F11119+3257. As above, the mass outflow rates bear large uncertainties and a large sample linking the small and large radii (as done for UFOs) is currently missing; we encourage observational proposals in such unification direction. We are living a new era for multiphase AGN outflows, as the field is rapidly growing via new high-resolution ALMA cycles able to probe $\sim 500 \text{ km s}^{-1}$ CO outflows (as shown by Morganti et al. 2015).

5. SUBGRID/SAM MODEL FOR AGN FEEDBACK

Below we describe how to incorporate our model into large-scale simulations of structure formation. Let us denote the typical resolution of a given simulation by Δr (nowadays $\sim 1 \text{ kpc} \gg r_s$ in a typical zoom-in run). Assuming the resolution is enough to resolve the thermalization region ($\Delta r < r_{\text{th}}$), we propose the following.

- (i) The SMBH growth can be tracked via Eq. 12 or 13,

$$\dot{M}_{\bullet} \simeq (0.03 \dot{M}_{\text{cool}}) T_{x,7.4} = (0.04 M_{\odot} \text{ yr}^{-1}) L_{x,43.8},$$

i.e., only a few percent of the macro cold inflow rate is actually deposited into the SMBH (with coarse resolution it may be easier to estimate the cold inflow from the core L_x , with the condition that the current central cooling time is shorter than $t_{\text{H}}/2$; see Eq. 7).

- (ii) The AGN mechanical feedback is injected on the scales defined by a few Δr with velocity given by Eq. 21,

$$v_{\text{OUT}} = \eta^{-1/2} v_{\text{out}},$$

where $\eta(r \approx \Delta r)$ is the entrainment factor at the resolved radial distance (Eq. 22) and v_{out} is the nuclear velocity of the outflow set by Eq. 17.

- (iii) The rate at which such outflow carries mass results from the entrainment mechanism given by Eq. 19,

$$\dot{M}_{\text{OUT}} = \eta \dot{M}_{\text{out}} \simeq \eta \dot{M}_{\text{cool}},$$

where \dot{M}_{cool} reflects the magnitude of the quenched cooling flow, which should self-consistently arise from the AGN feedback loop as a central cold inflow. The outflow can be injected as a mass flux through the boundary (e.g., sink the inflow rate and inject it back boosted by a factor η). If resolution does not permit to resolve the

⁸ In low-mass galaxies the thermalization radius is $< 10 \text{ kpc}$, thus the outflow can rapidly decline in velocity (and mass rate).

CCA inflow, it is better to not sink the gas and kick the gas mass per timestep over the most inner number of cells/particles (reaching \dot{M}_{OUT}) directly in the domain (checking for stability). Such inner active mass per timestep is naturally a fair representation of the entrained mass outflow rate (as tested in Gaspari et al. 2011b, 2012a). A remark is that for very coarse resolutions $\Delta r > r_{\text{th}}$ (Eq. 26), injecting massive outflows loses physical meaning, and the average radiative energy losses should be simply balanced via thermal energy injection, since the outflows are expected to be thermalized.

Such prescription is perfectly suited to be used also in semi-analytic models (SAM), e.g., of galaxy and cluster evolution, as well as in the interpretation of observational data (limited by the instrumental – instead of numerical – resolution). Furthermore, the injected properties, in particular the efficiency, are known a priori, regardless of numerics, implying that the fine-tuning loop plaguing current cosmological runs can be avoided (typically fitting one mass range, but overheating or overcooling the opposite regime due to keeping a constant macro efficiency). In other words, there is no main free parameter involved, except for the scatter intrinsic in observations. A sanity check is to retrieve the observed X-ray properties, e.g., X-ray luminosity and temperature profiles of the group or cluster. If not, the implementation of AGN feedback is numerically flawed and shall be modified accordingly, not retuning the parameters, but changing the injection implementation and carefully assessing which hydrodynamic solver and discretization to use. In other words, retuning some parameters to counteract the numerical flaws must be avoided, and can be avoided with the above a-priori prescription, thus preserving predictability.

6. DISCUSSION

We now discuss some details of the proposed model, together with the limitations and possible improvements.

The approach of this work differs from typical analytic modeling considering a perfect steady state solution (e.g., Bondi 1952) in which inflow and outflow coexist at exactly the same time (setting $\partial/\partial t = 0$ in the hydro equations). As indicated by X-ray observations and simulations (§1), the detailed self-regulated AGN feedback loop is time varying. We have instead considered a nearly stationary case over a feedback cycle, which is typically of the order of the central cooling time $t_{\text{cool}} \simeq k_{\text{b}}T/(n\Lambda)$, where Λ is the plasma cooling function (Sutherland & Dopita 1993); from isolated galaxies to massive clusters the typical central cooling time of the hot gas varies from tens to several 100 Myr (Gaspari et al. 2014). Within one cycle the process is time varying, with energy and mass changing form and phase. Specifically, the inflow acts first via the self-similar CCA rain, then the SMBH reacts to the feeding via nuclear ultra-fast outflows (Fig. 1). The propagating UFOs entrain the diffuse phase and thermalize in the core, such that $P_{\text{OUT}} \simeq L_{\text{x}}$, as shown by X-ray data (e.g., Main et al. 2017). The background halo is recurrently contracting and expanding in a gentle manner, and is never evacuated; in other words the core oscillates near HSE. Over the whole core region and one loop time the mass and energy are conserved (the small mass loss onto the BH is replenished from the virial hot halo). Note that if central $t_{\text{cool}} \gtrsim t_{\text{H}}/2$, the system is

in a non-cool-core condition and the feeding/feedback is not currently active. A key observational evidence for a variable feeding mechanism, is the ubiquitous variability of AGN light curves. As discussed in G17 (Sec. 5.1) and King & Nixon (2015), chaotic accretion drives a ‘flicker’ noise with major accretion events having Myr duration.

Needless to say a full, time-dependent treatment of the feeding and feedback process requires 3D (GR)MHD simulations covering the whole dynamical spatial and temporal range. However, until we will be able to break such computational barrier, we can rely on key properties of the inflows and outflows set by the multiwavelength constraints, which must be satisfied even in the advanced numerical runs. We remark X-ray data show that the feedback must be gentle and kinetically driven (with large-scale thermalization up to 100s kpc for massive clusters). Notice that the details of the energy conserving outflow are in our macro model not relevant. On the other hand, the momentum flux boost of the swept-up material due to the hot shocked gas and entrainment via hydro instabilities (e.g., Kelvin-Helmholtz and Rayleigh-Taylor) requires numerical simulations to be robustly understood. In addition to direct uplift, an interesting possibility to form molecular outflows is the in-situ condensation of the massive galaxy-scale hot wind via thermal instability – as discussed by Zubovas & King (2014) – which may further promote the subsequent precipitation phase.

In this work, we decided to aim for minimal assumptions and rely on first principles as much as possible. Further sophistications to the model are possible and can be easily incorporated to fit more specific objects, at the expense of an increased number of parameters. For instance, the inner background density profile can be modified with a more complex functional form than a single power law and/or assigning different volume filling profiles to the warm/cold phases. The configuration of the inner outflows can be modified by reducing Ω , in order to accommodate for a thinner bipolar setup. We note, in one loop, the cold inflow can occur along one direction, while the entrained outflow may occur in the perpendicular direction, further corroborating the separation of the large-scale CCA inflow and outflow mass rate, instead of a perfectly radial steady-state solution. A time delay in the loop can be introduced by tracking turbulent Taylor number: if $\text{Ta}_{\text{t}} > 1$, then a rotating structure (disc, ring, torus) can momentarily reduce accretion. We did not aim to fit one particular system or AGN outflow in this study, discussing only mean values. As noted in §4, considering the scatter in cooling system properties, the outflow variations are ~ 0.5 dex over a large sample. Fitting and interpreting single object data can be easily refined, e.g., by analyzing the core and nuclear X-ray spectrum both in terms of cooling rate (soft X-ray) and outflow line absorption features (hard X-ray).

Consistently with the observational results by Russell et al. 2013 (Fig. 12), the GR-RMHD simulations (SG17) show that for accretion rates below 10^{-2} the Eddington rate, the nuclear SMBH power is dominated by kinetic energy over the SMBH radiative output, $P_{\text{out}} \gg L_{\text{AGN}}$. The mechanical, sub-Eddington mode is the long-term maintenance mode of AGN feedback (McNamara & Nulsen 2012 for a review) preserving hot halos and cool-core systems in quasi thermal equilibrium at least for 9-

10 Gyr (McDonald et al. 2014, 2016, 2017). At high redshift ($z > 2$), the Eddington rate can be approached triggering a brief ‘quasar’ phase (seeding part of the SMBH mass). The wind may be thus radiatively driven, although its coupling with the gas is matter of ongoing debate. Moreover, there is no physical reason to think that the mechanical power from AGN is erased in this regime, as corroborated by our GR-RMHD run covering the quasar transition (see SG17). Even in such short-lived radiative regime, the outflow is still expected to be energy conserving⁹ (Faucher-Giguère & Quataert 2012) although it may be more appropriate to use a slightly larger $\varepsilon_{\bullet} \simeq 0.057$ (Novikov & Thorne 1973; Merloni & Heinz 2008) and rescale Eq. 9. As long as $\varepsilon_{\bullet} \gg \varepsilon_{\text{BH}}$, the outflow properties are however not significantly altered. A few quasar blasts may evacuate the system, but these anomalously powerful outbursts – which are much easier to detect – must be outliers (increasing the high-redshift population scatter) otherwise the majority of systems would later remain non-cool-core, which is not observed (e.g., Gaspari et al. 2014). Overall, regardless of the details of the driving mechanism (e.g., magnetic versus radiative), if self-regulation is on average preserved, the proposed model applies in similar way throughout cosmic time.

In the current interpretation, the micro and macro mechanical driver is a sub-relativistic outflow. Given the BH null spin expectation from chaotic accretion (King & Pringle 2006) and the high piercing collimation, a radio jet is expected to be subdominant, albeit it can coexist and trace the large-scale features, as bubbles. Observationally, radio synchrotron (electron) power is less than a percent of the cavity internal power (McNamara & Nulsen 2012), so only relativistic ions are left to inflate a bubble; however, this would produce strong Gamma emission in all systems, which *Fermi* does not typically observe. Moreover, several AGN bubbles are ghost cavities devoid of radio emission. Having said that, our model is general and the radio jet interpretation can be trivially implemented, e.g., by replacing the related micro efficiency and opening angle.

A current observational limitation which is worth discussing is the low-mass end regime. While hot, X-ray halos are well detected above stellar masses $M_* \gtrsim 10^{10.8} M_{\odot}$, in particular massive galaxies, galaxy groups and clusters, the precise level of the X-ray luminosity due to the diffuse component in the opposite regime ($T_x \lesssim 0.3 \text{ keV}$) is still uncertain due to the contamination of X-ray binaries (e.g., Anderson et al. 2015). The X-ray luminosity in such regime may be lower than our adopted scaling, and the relative cooling rate (Eq. 7) should be properly rescaled if necessary. While the outflow velocities are overall unaffected (Eq. 18-21), the mass outflow rate may be lower than the expected value. Conversely, while more massive systems have better constrained core X-ray luminosity, the stronger and harder diffuse emission substantially hinders the nuclear X-ray spectral features, making UFO detection challenging. If T_x is not available (e.g., for proto-galaxies), we suggest to use a core temperature in lower energy bands, as condensation occurs throughout the warm and cold phase regime.

⁹ As cooling acts on electrons, this slows down inverse Compton process; free-free cooling is secondary.

Finally, supernova feedback due to star formation (e.g., with rate a few percent of the galaxy cooling rate) can also become energetically important in low mass galaxies and shall be investigated in the future.

While here we have investigated the instantaneous properties as the SMBH accretion rates, $\dot{M}_{\bullet} \propto L_x$, in a separate work, we will focus on the integrated properties of the proposed unified model, as the total black hole masses and related scalings (e.g., the Magorrian relation). We anticipate some important considerations. As discussed above, the CCA self-regulation has a characteristic frequency related to the cooling time, $1/t_{\text{cool}}$, as the hot halo requires such time to promote condensation, rain down, and then activate the ultra-fast outflow feedback. One loop requires $t_{\text{cyc}} = t_{\text{cool}} + t_{\text{OUT}} \approx t_{\text{cool}}$ (the outflow active time is always shorter than the condensation time). In other words, the duty cycle increases from clusters to galaxies, as corroborated by long-term AGN feedback simulations (e.g., Gaspari et al. 2011a,b, 2012a) and X-ray shocks/cavities observations (e.g., Randall et al. 2015). The number of cycles over the Hubble time is thus $n_{\text{cyc}} = t_{\text{H}}/t_{\text{cool}}$, with an active time $t_{\text{act}} = n_{\text{cyc}} t_{\text{OUT}}$. The black hole masses are expected to grow as $M_{\bullet} \simeq \dot{M}_{\bullet} t_{\text{act}}$, hence with a temperature scaling given by $M_{\bullet} \propto L_x/t_{\text{cool}} \propto T_x^3/(T_x/\Lambda) \propto T_x^2 \propto \sigma_*^4$, as core temperature is a measure of the (stellar) velocity variance in virialized structures. This is valid in the galactic regime ($T_x \approx 0.5 - 2 \text{ keV}$) as Λ remains essentially constant for solar metallicity. For clusters, $\Lambda \propto T^{1/2}$ due to Bremsstrahlung, thus $M_{\bullet} \propto T_x^{2.5} \propto \sigma_*^5$. Observations show a very similar scaling, with ultramassive black holes found predominantly in more massive halos which are consistent with our self-regulated CCA model inducing a steepening of the Magorrian relation (e.g., Gültekin et al. 2009; McConnell & Ma 2013; Kormendy & Ho 2013).

7. SUMMARY AND CONCLUSIONS

We linked for the first time the *physical* micro and macro mechanical efficiency of SMBHs, the latter based on key X-ray data and hydrodynamical simulations, the former retrieved by state-of-the-art GR-RMHD horizon simulations, such that $\varepsilon_{\text{BH}} = 10^{-3} T_{x,7.4}$ and $\varepsilon_{\bullet} = 0.03$, respectively (§2). By using minimally first principles, as conservation of energy ($P_{\text{out}} = P_{\text{OUT}} \simeq L_x$, where the latter is the core luminosity of the hot halo), we unified the macro and micro properties of self-regulated AGN feedback from the galactic to the cluster regime (§3).

The inflow mechanism occurs via chaotic cold accretion (CCA) – probed during the last years – i.e., the rain of cold clouds condensing out of the quenched cooling flow (\dot{M}_{cool}), which are recurrently funneled via fractal, inelastic collisions. Near hundreds gravitational radii, the binding energy of accreting gas is strongly transformed into ultrafast outflows (UFOs) with characteristic velocity of a few 10^4 km s^{-1} ($\sqrt{2\varepsilon_{\text{BH}}c}$) ejecting most of the inflowing gas mass as $\dot{M}_{\text{out}} \approx \dot{M}_{\text{cool}}$ ($\approx 1 M_{\odot} \text{ yr}^{-1}$ for intermediate systems).

At larger radii, the outflow entrains progressively more mass, such as $\dot{M}_{\text{OUT}} = \eta \dot{M}_{\text{out}}$ and $v_{\text{OUT}} = \eta^{-1/2} v_{\text{out}}$, with $\eta \propto r^{2/3}$. At roughly the kpc scale, the characteristic velocities of large-scale hot/warm/cold outflows are predicted to be a few 10^3 , 1000, and 500 km s^{-1} , re-

spectively (depending on the inner dominant gas phase). The related average mass outflow rates (for 1 keV systems) are expected to be of the order of 10, 100, several $100 M_{\odot} \text{ yr}^{-1}$, respectively. Such properties are in agreement with observations of UFOs, and kpc-scale ionized, neutral, and molecular outflows (§4). Velocities are the more robust and stable indicator compared with outflow rates, both observationally and in the model. Ultimately, the outflows thermalize within the system core ($\lesssim 0.1 r_{\text{vir}}$), balancing the cooling losses, and allowing another self-regulated loop to reload via CCA rain and outflow feedback – with frequency $\propto t_{\text{cool}}^{-1}$.

A key aspect of the newly presented model is that the irradiated cool-core energy rate (L_x) reflects the gas flow onto the tiny SMBH, creating a symbiotic link over a 10 dex dynamical range. The tiny SMBHs are not isolated point objects where space-time diverges, but appear to be central actors in the evolution of both the micro and cosmic structures. In particular, the SMBH growth rate is linked to the large-scale T_x halo and thus any other cosmic scaling (e.g., L_x, M_{vir}), in addition to inducing a consistent $M_{\bullet} - \sigma_*$ relation. Despite the necessary limitations (§6), the CCA+UFO model captures the essential ingredients than any more sophisticated self-regulation model and simulation should have at its core, in particular the gentle quasi-thermal equilibrium of plasma halos.

The pursued minimalism of the CCA+UFO model makes it suited to be trivially implemented in subgrid modules and semi-analytic works (§5), as well as in estimates for the interpretation of observational studies, e.g., related to nuclear and entrained outflow velocities and mass rates. The proposed model presents a simple *physical* unification scheme upon which construct and conduct future multiwavelength investigations, e.g., selecting the systems in terms of the core X-ray luminosity (or other related macro observable). Instead of classifying a phenomenological aspect of a peculiar AGN, we encourage observational campaigns in the direction of understanding the common, unified physics of multiphase inflows/outflows (e.g., §4) and to systematically consider the connection between the AGN and the global hot halo. A larger and homogeneous X-ray, optical, and radio sample of such properties, from low-mass galaxies to massive clusters, is needed to robustly test the link of the micro and macro properties of AGN feedback.

ACKNOWLEDGEMENTS

MG and AS acknowledge support for this work by NASA through Einstein Postdoctoral Fellowship number PF5-160137 and PF4-150126 awarded by the Chandra X-ray Center, which is operated by the Smithsonian Astrophysical Observatory for NASA under contract NAS8-03060. Support for this work was also provided by NASA Chandra award number G07-18121X. FLASH code was in part developed by the DOE NNSA-ASC OASCR Flash center at the University of Chicago. HPC resources were provided by the PL-Grid Infrastructure and the NASA/Ames HEC Program (SMD-16-7251). We thank B. McNamara, G. Tremblay, J. Stone, M. McDonald, R. Morganti, F. Tombesi, M. Cappi, and F. Combes for insightful discussions.

REFERENCES

- Anderson M. E., Gaspari M., White S. D. M., Wang W., Dai X., 2015, *MNRAS*, **449**, 3806
- Bondi H., 1952, *MNRAS*, **112**, 195
- Cicone C., et al., 2014, *A&A*, **562**, A21
- Combes F., 2015, in Ziegler B. L., Combes F., Dannerbauer H., Verdugo M., eds, IAU Symposium Vol. 309, Galaxies in 3D across the Universe. pp 182–189 ([arXiv:1408.1591](https://arxiv.org/abs/1408.1591)), [doi:10.1017/S1743921314009636](https://doi.org/10.1017/S1743921314009636)
- Crenshaw D. M., Kraemer S. B., George I. M., 2003, *ARA&A*, **41**, 117
- David L. P., et al., 2014, *ApJ*, **792**, 94
- Ettori S., Fabian A. C., 2000, *MNRAS*, **317**, L57
- Faucher-Giguère C.-A., Quataert E., 2012, *MNRAS*, **425**, 605
- Feruglio C., Maiolino R., Piconcelli E., Menci N., Aussel H., Lamastra A., Fiore F., 2010, *A&A*, **518**, L155
- Feruglio C., et al., 2015, *A&A*, **583**, A99
- Fukumura K., Kazanas D., Contopoulos I., Behar E., 2010, *ApJ*, **715**, 636
- Fukumura K., Tombesi F., Kazanas D., Shrader C., Behar E., Contopoulos I., 2014, *ApJ*, **780**, 120
- Fukumura K., Tombesi F., Kazanas D., Shrader C., Behar E., Contopoulos I., 2015, *ApJ*, **805**, 17
- Gaspari M., 2015, *MNRAS*, **451**, L60
- Gaspari M., Melioli C., Brighenti F., D’Ercole A., 2011a, *MNRAS*, **411**, 349
- Gaspari M., Brighenti F., D’Ercole A., Melioli C., 2011b, *MNRAS*, **415**, 1549
- Gaspari M., Brighenti F., Temi P., 2012a, *MNRAS*, **424**, 190
- Gaspari M., Ruszkowski M., Sharma P., 2012b, *ApJ*, **746**, 94
- Gaspari M., Ruszkowski M., Oh S. P., 2013, *MNRAS*, **432**, 3401
- Gaspari M., Brighenti F., Temi P., Ettori S., 2014, *ApJ*, **783**, L10
- Gaspari M., Brighenti F., Temi P., 2015, *A&A*, **579**, A62
- Gaspari M., Temi P., Brighenti F., 2017, *MNRAS*, **466**, 677
- Gültekin K., et al., 2009, *ApJ*, **698**, 198
- Heckman T. M., Best P. N., 2014, *ARA&A*, **52**, 589
- Kaastra J. S., et al., 2004, *A&A*, **413**, 415
- Khatri R., Gaspari M., 2016, *MNRAS*, **463**, 655
- King A., Nixon C., 2015, *MNRAS*, **453**, L46
- King A. R., Pounds K. A., 2014, *MNRAS*, **437**, L81
- King A. R., Pringle J. E., 2006, *MNRAS*, **373**, L90
- Kormendy J., Ho L. C., 2013, *ARA&A*, **51**, 511
- Kravtsov A. V., Borgani S., 2012, *ARA&A*, **50**, 353
- Main R. A., McNamara B. R., Nulsen P. E. J., Russell H. R., Vantyghem A. N., 2017, *MNRAS*, **464**, 4360
- McConnell N. J., Ma C.-P., 2013, *ApJ*, **764**, 184
- McDonald M., Roediger J., Veilleux S., Ehlert S., 2014, *ApJ*, **791**, L30
- McDonald M., et al., 2016, *ApJ*, **826**, 124
- McDonald M., et al., 2017, preprint, ([arXiv:1702.05094](https://arxiv.org/abs/1702.05094))
- McNamara B. R., Nulsen P. E. J., 2012, *New J. Phys.*, **14**, 055023
- McNamara B. R., Russell H. R., Nulsen P. E. J., Hogan M. T., Fabian A. C., Pulido F., Edge A. C., 2016, *ApJ*, **830**, 79
- Merloni A., Heinz S., 2008, *MNRAS*, **388**, 1011
- Moller A., Sadowski A., 2015, preprint, ([arXiv:1509.06644](https://arxiv.org/abs/1509.06644))
- Morganti R., 2015, in Massaro F., Cheung C. C., Lopez E., Siemiginowska A., eds, IAU Symposium Vol. 313, Extragalactic Jets from Every Angle. pp 283–288 ([arXiv:1411.6107](https://arxiv.org/abs/1411.6107)), [doi:10.1017/S1743921315002331](https://doi.org/10.1017/S1743921315002331)
- Morganti R., Tadhunter C. N., Oosterloo T. A., 2005, *A&A*, **444**, L9
- Morganti R., Holt J., Saripalli L., Oosterloo T. A., Tadhunter C. N., 2007, *A&A*, **476**, 735
- Morganti R., Oosterloo T., Oonk J. B. R., Frieswijk W., Tadhunter C., 2015, *A&A*, **580**, A1
- Narayan R., Yi I., 1995, *ApJ*, **452**, 710
- Nesvadba N. P. H., et al., 2010, *A&A*, **521**, A65
- Novikov I. D., Thorne K. S., 1973, in Dewitt C., Dewitt B. S., eds, Black Holes (Les Astres Occlus). pp 343–450
- Peterson J. R., Kahn S. M., Paerels F. B. S., Kaastra J. S., Tamura T., Bleeker J. M., Ferrigno C., Jernigan J., 2003, *ApJ*, **590**, 207
- Pizzolato F., Soker N., 2010, *MNRAS*, **408**, 961
- Prasad D., Sharma P., Babul A., 2015, *ApJ*, **811**, 108
- Randall S. W., et al., 2015, *ApJ*, **805**, 112
- Russell H. R., McNamara B. R., Edge A. C., Hogan M. T., Main R. A., Vantyghem A. N., 2013, *MNRAS*, **432**, 530
- Russell H. R., et al., 2014, *ApJ*, **784**, 78

- Russell H. R., et al., 2016, preprint, ([arXiv:1611.00017](https://arxiv.org/abs/1611.00017))
- Sadowski A., Gaspari M., 2017, preprint, ([arXiv:1701.07033](https://arxiv.org/abs/1701.07033))
- Sądowski A., Narayan R., Tchekhovskoy A., Abarca D., Zhu Y., McKinney J. C., 2015, *MNRAS*, **447**, 49
- Sądowski A., Lasota J.-P., Abramowicz M. A., Narayan R., 2016, *MNRAS*, **456**, 3915
- Soker N., 2016, *New A Rev.*, **75**, 1
- Soker N., Sternberg A., Pizzolato F., 2009, in Heinz S., Wilcott E., eds, American Institute of Physics Conference Series Vol. 1201, American Institute of Physics Conference Series. pp 321–325 ([arXiv:0909.0220](https://arxiv.org/abs/0909.0220)), [doi:10.1063/1.3293066](https://doi.org/10.1063/1.3293066)
- Sturm E., et al., 2011, *ApJ*, **733**, L16
- Sun M., 2012, *New Journal of Physics*, **14**, 045004
- Sun M., Voit G. M., Donahue M., Jones C., Forman W., Vikhlinin A., 2009, *ApJ*, **693**, 1142
- Sutherland R. S., Dopita M. A., 1993, *ApJS*, **88**, 253
- Tchekhovskoy A., Narayan R., McKinney J. C., 2011, *MNRAS*, **418**, L79
- Teng S. H., Veilleux S., Baker A. J., 2013, *ApJ*, **765**, 95
- Tombesi F., 2016, *Astronomische Nachrichten*, **337**, 410
- Tombesi F., Cappi M., Reeves J. N., Braito V., 2012, *MNRAS*, **p. L413**
- Tombesi F., Cappi M., Reeves J. N., Nemmen R. S., Braito V., Gaspari M., Reynolds C. S., 2013, *MNRAS*, **430**, 1102
- Tombesi F., Tazaki F., Mushotzky R. F., Ueda Y., Cappi M., Gofford J., Reeves J. N., Guainazzi M., 2014, *MNRAS*, **443**, 2154
- Tombesi F., Meléndez M., Veilleux S., Reeves J. N., González-Alfonso E., Reynolds C. S., 2015, *Nature*, **519**, 436
- Tremblay G. R., et al., 2016, *Nature*, **534**, 218
- Vikhlinin A., Kravtsov A., Forman W., Jones C., Markevitch M., Murray S. S., Van Speybroeck L., 2006, *ApJ*, **640**, 691
- Voit G. M., Donahue M., Bryan G. L., McDonald M., 2015, *Nature*, **519**, 203
- Werner N., et al., 2014, *MNRAS*, **439**, 2291
- Yang H.-Y. K., Reynolds C. S., 2016, *ApJ*, **829**, 90
- Zubovas K., King A. R., 2014, *MNRAS*, **439**, 400

APPENDIX

CORE LUMINOSITY AND TEMPERATURE

Most of the X-ray luminosity comes from the region well within r_{500} due to the steep radial density profile (emissivity is $\propto \rho(r)^2$). By using the available *Chandra* and *XMM* (losing sensitivity at large radii) luminosities is thus a fair proxy for the core luminosity. More accurately, we can model the surface brightness with a β profile, $SB_x = SB_0 (1 + R^2/r_c^2)^{-3\beta+1/2}$, where R is the projected radius and SB_0 is the inner normalization. Integrating over thin annuli yields

$$L_x(< r) = SB_0 \frac{2\pi r_c^2}{3(2\beta - 1)} \left[1 - \left(1 + \frac{R^2}{r_c^2} \right)^{-3\beta+3/2} \right]. \quad (\text{A1})$$

The cooling radius is typically equal to the core radius ($\approx 0.2 r_{500}$; Vikhlinin et al. 2006), since the radial breaking naturally emerges via the loss of pressure, $r_{\text{cool}} \simeq r_c$ (Ettori & Fabian 2000). Cool-core systems are better fitted by a sum of two beta models for the core and the outskirt; characteristic values are $\beta_c \approx 1.7$ and $\beta_o \approx 0.7$, respectively (e.g., Ettori & Fabian 2000). Plugging in this values in a double β model following each Eq. A1, the average correction for the core luminosity is 0.68 L_{500} . Notice that the outskirts, $r_{\text{vir}} \simeq 2 r_{500}$, contribute in negligible measure, $L_{\text{vir}}/L_{500} \approx 1.05$. Overall, the chosen luminosity radius does not significantly alter the results presented in §3. The temperature profile shows even less variation than density, varying by a factor 2-3. By emission-weighting it, the core T_x is typically 10 percent lower than the ambient T_{500} (Ettori & Fabian 2000; Vikhlinin et al. 2006) – again, a minor variation.

For an idealized self-similar spherical collapse, it is well known that $L_x \propto T_x^2$ (Kravtsov & Borgani 2012). However, observational data show that non-gravitational/feedback processes steepen such relation as $L_{500} \simeq 8.8 \times 10^{43} (T_{500}/2.5 \text{ keV})^3$ (Sun 2012). In §3, we are interested in the X-ray luminosity and temperature tied to the core/cooling region, i.e., the radius within which the temperature profile slope becomes positive ($r_c \approx 0.2 r_{500}$, related to $t_{\text{cool}} \approx t_{\text{age}} \sim t_H/2$, where t_H is the Hubble time). By using the above minor corrections, the core scaling relation becomes $L_x \simeq 6 \times 10^{43} (T_x/2.2 \text{ keV})^3 \text{ erg s}^{-1}$. For reference, in the local universe, the scaling between radius and temperature ($r^3 \propto M \propto T^{3/2}$) is $r_{500} \simeq (0.74 \text{ Mpc}) T_{x,7.4}^{1/2}$ (Sun et al. 2009), leading to a physical core radius $r_c \approx (148 \text{ kpc}) T_{x,7.4}^{1/2}$.