. SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI

SISSA Digital Library

Post-Newtonian evolution of massive black hole triplets in galactic nuclei - III. A robust lower limit to the nHz stochastic background of gravitational waves

Original
Post-Newtonian evolution of massive black hole triplets in galactic nuclei - III. A robust lower limit to the nHz stochastic background of gravitational waves / Bonetti, M., Sesana, A., Barausse, E., Haardt, F.. - In: MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY. - ISSN 0035-8711. - 477:2(2018), pp. 2599-2612. [10.1093/mnras/sty874]

Availability:
This version is available at: 20.500.11767/89654 since: 2019-04-16T12:41:25Z

Publisher:

Published
DOI:10.1093/mnras/sty874

Terms of use:

Testo definito dall'ateneo relativo alle clausole di concessione d'uso

Publisher copyright
Oxford University Press

This version is available for education and non-commercial purposes.

note finali coverpage

(Article begins on next page)

11 September 2026

Mon. Not. R. Astron. Soc. **000**, 1–?? (2017) Printed 10 April 2018 (MN LATEX style file v2.2)

Post-Newtonian evolution of massive black hole triplets in galactic nuclei – III. A robust lower limit to the nHz stochastic background of gravitational waves.

Matteo Bonetti[1,2], Alberto Sesana[3], Enrico Barausse[4,5] & Francesco Haardt[1,2]
[1] *DiSAT, Università degli Studi dell'Insubria, Via Valleggio 11, 22100 Como, Italy*
[2] *INFN, Sezione di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy*
[3] *Institute of Gravitational Wave Astronomy and School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom*
[4] *CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis Bd Arago, 75014 Paris, France*
[5] *Sorbonne Universités, UPMC Univesité Paris 6, UMR 7095, Institut d'Astrophysique de Paris, 98 bis Bd Arago, 75014 Paris, France*

ABSTRACT
Inspiraling massive black-hole binaries (MBHBs) forming in the aftermath of galaxy mergers are expected to be the loudest gravitational-wave (GW) sources relevant for pulsar-timing arrays (PTAs) at nHz frequencies. The incoherent overlap of signals from a cosmic population of MBHBs gives rise to a stochastic GW background (GWB) with characteristic strain around $h_c \sim 10^{-15}$ at a reference frequency of 1 yr^{-1}, although uncertainties around this value are large. Current PTAs are piercing into the GW amplitude range predicted by MBHB-population models, but no detection has been reported so far. To assess the future success prospects of PTA experiments, it is therefore important to estimate the minimum GWB level consistent with our current understanding of the formation and evolution of galaxies and massive black holes (MBHs). To this purpose, we couple a semianalytic model of galaxy evolution and an extensive study of the statistical outcome of triple MBH interactions. We show that even in the most pessimistic scenario where all MBHBs stall before entering the GW-dominated regime, triple interactions resulting from subsequent galaxy mergers inevitably drive a considerable fraction of the MBHB population to coalescence. At frequencies relevant for PTA, the resulting GWB is only a factor of 2-to-3 suppressed compared to a fiducial model where binaries are allowed to merge over Gyr timescales. Coupled with current estimates of the expected GWB amplitude range, our findings suggest that the minimum GWB from cosmic MBHBs is unlikely to be lower than $h_c \sim 10^{-16}$ (at $f = 1$ yr^{-1}), well within the expected sensitivity of projected PTAs based on future observations with FAST, MeerKAT and SKA.

Key words: black hole physics – galaxies: kinematics and dynamics – gravitation – gravitational waves – methods: numerical

1 INTRODUCTION

Massive black holes (MBHs) are among the primary building blocks of galaxies. Since the early nineties it has become clear that most (if not all) massive galaxies host a MBH at their centre (Kormendy & Richstone 1995) whose mass correlates with the properties of the galactic host (Magorrian et al. 1998; Ferrarese & Merritt 2000; Gebhardt et al. 2000), pointing towards a co-evolution of MBHs and their host galaxies. In the standard hierarchical framework of structure formation, galaxies form in a bottom-up fashion, whereby the massive galaxies that we see today build up at the intersection of dark matter filaments along which galaxies and cold gas can stream inwards without getting shock-heated (Dekel, Sari & Ceverino 2009). At those intersections, galaxies experience a sequence of mergers and accretion events, that contribute to their final mass. If MBHs are ubiquitous in galaxy centres, then MBH binaries (MBHBs) naturally form following galaxy mergers (Begelman, Blandford & Rees 1980). If MBHs can efficiently pair into sub-parsec binaries, then gravitational wave (GW) emission inevitably takes over (Peters & Mathews 1963), leading to final coalescence. In this scenario, MBHs grow through a series of mergers and accretion events, and the interplay of accretion and feedback is thought to be responsible of the observed MBH-galaxy relations (Kauffmann & Haehnelt 2000; Volonteri, Haardt & Madau 2003).

MBHBs are the loudest GW sources in the Universe, and a typical billion-solar-mass system inspiralling at centi-parsec scales emits GWs at nHz frequencies (Sesana, Vecchio & Colacino 2008), currently probed by pulsar-timing arrays (PTAs, Foster & Backer 1990). In fact, milli-second pulsars (MSPs) are the most stable macroscopic clocks known in Nature (Taylor 1992), and a GW passing through their line of sight leaves a characteristic imprint on the time of arrivals (TOAs) of the pulses on Earth (Sazhin 1978). By working within the hierarchical structure-formation paradigm, it is possible to construct the expected distribution of MBHBs populating the Universe, as a function of mass, redshift and orbital frequency. The GW signals from individual binaries will add up incoherently to form a stochastic GW background (GWB, Rajagopal & Romani 1995; Jaffe & Backer 2003; Wyithe & Loeb 2003; Sesana et al. 2004), which in its simplest form (i.e. assuming a population of circular, GW driven MBHBs) has a single power-law spectrum with characteristic strain parametrised as $h_c = A[f/(1\ \mathrm{yr}^{-1})]^{-2/3}$ (Phinney 2001), where f is the observed frequency. Note that essentially all results in the PTA literature are quoted as strain at $f = 1\ \mathrm{yr}^{-1}$, we therefore quote our bulk results in terms of the normalisation A.

The amplitude of the signal has been computed by several authors in the past two decades using several approaches, including: analytic and semianalytic models for structure formation and the MBHB merger rate (Wyithe & Loeb 2003; Sesana, Vecchio & Colacino 2008; McWilliams, Ostriker & Pretorius 2014), empirical galaxy merger rate estimates from observations (Rajagopal & Romani 1995; Jaffe & Backer 2003; Sesana 2013; Ravi et al. 2015; Sesana et al. 2016) and large cosmological simulation (Sesana, Vecchio & Volonteri 2009; Ravi et al. 2012; Kulier et al. 2015; Kelley et al. 2017). The typical value of A inferred from these theoretical models is of the order of $\sim 10^{-15}$. This is particularly interesting, because 10^{-15} is the sensitivity that is achievable by timing about ten MSPs with a precision of about 100 ns over several years (Jenet et al. 2006), which is now within reach of the leading PTA experiments. In fact the European PTA (EPTA, Desvignes et al. 2016), NANOGrav (The NANOGrav Collaboration et al. 2015), the Parkes PTA (PPTA, Reardon et al. 2016) and the International PTA (IPTA, Hobbs et al. 2010) have recently placed upper limits on the amplitude of a stochastic GWB at a level of $A = 3\times10^{-15}, 1.5\times10^{-15}, 1\times10^{-15}, 1.7\times10^{-15}$ respectively (Lentati et al. 2015; Arzoumanian et al. 2016; Shannon et al. 2015; Verbiest et al. 2016).

PTAs are becoming a world-wide effort that requires the investment of significant human resources and radio-telescope time, including upcoming facilities like FAST, MeerKAT and, eventually, the SKA (see Janssen et al. 2015, for future prospects). It is therefore of paramount importance to carry out a comprehensive exploration of the predicted GW signal in order to critically assess the chances of success of this endeavour. The aforementioned theoretical models considered a wide range of scenarios, differing in several physical ingredients, including: i) a range of galaxy-evolution models and underlying galaxy-merger rates; ii) several MBH-galaxy scaling relations; iii) the role of gas and stellar dynamics in driving the MBHB at low frequencies; iv) eccentricity. In general, however, there is a common feature to all those models: MBHBs can efficiently reach the centi-parsec separations relevant for PTA observations (notably, Kelley et al. 2017, explored a range of efficiencies in the context of stellar driven MBHBs). Although this is a reasonable assumption, MBHs have to undergo a long journey before getting to those small separations. In fact, following galaxy mergers, dynamical friction is only efficient in driving the two MBHs to form a Keplerian binary (Begelman, Blandford & Rees 1980). For billion-solar-mass systems, this happens at separations of the order of tens of parsecs, where GW emission is still inefficient. Hardening driven by interactions with the dense stellar environment of galactic nuclei (e.g., Khan et al. 2012; Sesana & Khan 2015; Vasiliev, Antonini & Merritt 2015) or by a putative circumbinary disc (e.g., MacFadyen & Milosavljević 2008; Cuadra et al. 2009; Nixon et al. 2011) can bridge the gap, taking the MBHBs down to centi-pc separation. However, the efficiency of these processes critically depends on a number of physical conditions, such as the stellar density in the nuclei of very massive galaxies, the efficiency of relaxation processes in bringing stars on almost radial orbits that intersect the MBHB, or simply the mere availability of cold gas to form a sizeable circumbinary disc. This is sometimes referred to in the literature as the "final-parsec problem" (see Dotti, Sesana & Decarli 2012, for an overview of general issues related to bound MBHB evolution).

It is therefore of great interest to ask the following question: what if all hardening mechanisms fail and MBHBs simply stall? Would that be the tombstone of PTA experiments? This scenario has recently been explored by Dvorkin & Barausse (2017) (hereinafter DB17); they argued that even if all binaries stall, there would be a (much reduced) leftover GW signal in the PTA band. This is because the typical MBHB stalling radius is a function of the binary mass and mass ratio, and even though billion-solar-mass systems stall well outside the PTA band, lighter MBHB and low mass-ratio ones can still emit some GWs at nHz frequencies. The signal they predict is, however, at a level of $A \lesssim 10^{-16}$. A crucial ingredient that was missing in the DB17 modelling, however, is the formation of multiple MBH systems. In fact, massive galaxies typically experience multiple mergers along their formation history (see, e.g., Rodriguez-Gomez et al. 2015). If a MBHB stalls, the subsequent merger event will bring a third MBH, thus forming a MBH triplet (Volonteri, Haardt & Madau 2003; Iwasawa, Funato & Makino 2006; Hoffman & Loeb 2007; Kulkarni & Loeb 2012; Bonetti et al. 2016, Paper I hereinafter).

In this paper we explore the effect of MBH triple and quadruple interaction on the GW signal generated by the cosmic MBH population in the PTA band[1]. We model the statistics of MBH binaries, triplets and quadruplets by coupling the large suite of simulations described in Bonetti et al. (2017) (Paper II hereinafter) to the semianalytic model of galaxy formation of Barausse (2012). We consider two conceptually different situations: a fiducial model in which MBHBs merge on timescales of millions-to-billions year (depending on galactic properties) after their host galaxies

[1] While completing this draft, we became aware of a similar independent investigation by Ryu and collaborators (Ryu et al. 2017). Albeit employing different frameworks, the two studies reach similar conclusions

merge, and an extreme model in which all MBHBs stall at about their hardening radius (see Paper II for details), and mergers are prompted only by multiple MBH interactions. Within the context of the assumed galaxy-formation model, the latter is the most pessimistic scenario from the point of view of the expected GW signal in the PTA band. The difference between the two scenarios is indicative of the typical maximum suppression of the GW signal, for a given galaxy formation model, due to MBHB stalling. We show that multiple MBH interactions prompt the coalescence of a number of MBHBs which results, in the relevant PTA frequency band, in a GWB only a factor 2-to-3 reduced with respect to the efficient binary merger case. This is to be compared to the suppression factor of 10 or larger found by DB17, which neglected multiple MBH interactions. Our results imply that even a combination of a particularly unfavourable MBH-galaxy scaling relations (Sesana et al. 2016) and MBHB stalling still results in a GWB at the $A \approx 10^{-16}$ level, well within the capabilities of a realistic SKA-era PTA.

The paper is organised as follows. In Section 2 we describe the MBH evolution model and the implementation of triple and quadruple interactions. Section 3 summarises the method for the GWB computation. Our main results are presented in Section 4, and a few important caveats are discussed in Section 5. We conclude with some final remarks and directions for future investigation in Section 6. We assume a concordance Λ–CDM universe with $\Omega_M = 0.3$, $\Omega_\Lambda = 0.7$, and $H_0 = 70$ km/(s Mpc). Unless otherwise specified, we use geometric units where $G = c = 1$.

2 SEMIANALYTIC MODEL OF GALAXY AND MASSIVE BLACK HOLE EVOLUTION

2.1 General description of the model

We simulate the co-evolution of MBHs and their host galaxies by the semianalytic galaxy-formation model of Barausse (2012), with later incremental improvements described in Sesana et al. (2014); Antonini, Barausse & Silk (2015a,b). We refer to those references for a detailed description of the model. Here, we limit ourselves to summarising its most salient features. The model's calibration is described in Barausse et al. (2017), and reproduces the conservative MBH scaling relations of Shankar et al. (2016), which are known to produce a low level of the stochastic GW signal for PTAs (Sesana et al. 2016).

The evolution of Dark-Matter halos is modelled via merger trees produced with an extended Press-Schechter formalism, suitably modified so as to reproduce the results of N-body simulations (Press & Schechter 1974; Parkinson, Cole & Helly 2008). The baryonic components of galaxies are then evolved along the branches of these merger trees, while the nodes of the trees correspond to the moment when two halos touch, thus initiating the processes leading to halo, galaxy and eventually black-hole coalescence.

In more detail, galaxies form from either the cooling of an unprocessed "hot" gas component shock-heated to the halo's virial temperature, or (especially in low-mass systems and at high redshift) from accretion flows of colder gas (Dekel & Birnboim 2006; Cattaneo et al. 2006; Dekel et al. 2009). When the gas has reached the halo's centre as a result of either of these channels, it forms a disc, simply by conservation of angular momentum, and eventually starts forming stars. Galactic spheroids form instead when the gaseous and galactic discs are destroyed either by bar instabilities or by major galactic mergers. Both these processes are also assumed to drive cold gas to the galactic centre, thus enhancing star formation. During star formation episodes (in both discs and spheroids) we also account for the feedback from supernova explosions on the surrounding gas.

MBHs are formed from high-redshift seeds, with several possible plausible choices for their initial mass function and halo occupation fraction. In this work, we consider a "light-seed" (LS) scenario, where seeds of a few hundred $M_\odot$ are provided by the remnants of popIII stars forming in low-metallicity high-redshift galaxies (Madau & Rees 2001), and a "heavy-seed" (HS) scenario where larger ($\sim 10^5 M_\odot$) seeds form from the collapse (e.g. due to bar instabilities) of protogalactic discs. More precisely, in the LS scenario we populate with seeds only the most massive halos (i.e. those collapsing from the 3.5σ peaks of the primordial density field) at redshift $15 < z < 20$. The mass of each seed is assumed to be $\sim 2/3$ of the initial popIII star's mass (to account for mass losses during the collapse of the star). The mass of the star is drawn randomly from a log-normal distribution centred at 300 $M_\odot$ with rms of 0.2 dex and an exclusion region between 140 and 260 $M_\odot$, since stars in this range explode as pair-instability supernovae without forming black holes (Heger & Woosley 2002). For the HS scenario, we follow Volonteri, Lodato & Natarajan (2008), which models the formation of seeds from disc bar instabilities at redshift $15 < z < 20$. The model has just one free parameter, i.e. the critical Toomre parameter Q_c at which the instability sets in. The most likely values for Q_c range between 2 and 3. Here, we adopt $Q_c = 2.5$.

The black-hole seeds then grow via accretion and mergers. For the former channel, we assume that a gas reservoir forms in the nuclear region of each galaxy as a result of cold gas being funnelled to the centre during major galactic mergers or bar instabilities of the gaseous galactic discs. Because both kinds of events are also thought to trigger spheroid formation, we follow Granato et al. (2004); Lapi et al. (2014) and assume that the feeding of this nuclear reservoir is linearly correlated with the star formation rate in the spheroid component. The MBH then accretes from this reservoir on the viscous timescale, but we cap this accretion rate at the Eddington rate in the HS scenario, and at twice the Eddington rate in the LS one. [2] As a result, MBHs undergo periods of quiescent activity interrupted by quasar/AGN phases. The feedback of the MBH on the surrounding gas is taken into account in both phases (radio-mode and quasar feedback). The nuclear reservoir is also assumed to form stars, which give rise to a nuclear star cluster (Antonini, Barausse & Silk 2015a,b). We also assume that nuclear star clusters form and grown via a second channel, namely the migra-

[2] This choice is made because a certain degree of super-Eddington accretion is needed in an LS scenario, if one wants to reproduce the high-redshift AGN luminosity function, c.f. Madau, Haardt & Dotti (2014).

tion of globular clusters to the nuclear region induced by dynamical friction (Antonini, Barausse & Silk 2015a,b).

After two halos start coalescing (at the nodes of the merger tree), the smaller halo (the "satellite") initially retains its identity within the bigger one (the "host"), slowly falling towards the centre driven by dynamical friction. We account for this phase by using the expression for the dynamical-friction time from Boylan-Kolchin, Ma & Quataert (2008), which is calibrated against numerical simulations and accounts for the effect of both Dark Matter and baryons. We also model the mass loss incurred by the satellite halo and its galaxy due to tidal stripping and evaporation (Taffoni et al. 2003). When the satellite subhalo and its galaxy merge with the host, the MBHs contained in the two galaxies may still be very far apart (at $\sim$ kpc distances). Nevertheless, at least when the satellite and the host have mass ratios $\gtrsim 0.1$ (the most relevant case for our results, c.f. section 5), dynamical friction against the gas and stars of the newly formed galaxy quickly drives the MBHs toward the centre (Dosopoulou & Antonini 2017, DB17,). This process is particularly efficient because at least in its initial stages, the MBHs are expected to be still surrounded by a stellar core from their host galaxy, which results in a shorter dynamical-friction timescale. Therefore, for the purpose of this work, we can safely neglect this phase and assume that whenever the host and satellite galaxies coalesce, the MBHs are efficiently driven down to a separation comparable to the primary MBH influence radius $r_i \approx 2Gm_1/\sigma^2$ (m_1 being the primary black hole's mass and σ the velocity dispersion of the host's spheroid).

We then account for the subsequent evolution of the MBHs with the simple prescriptions outlined in (Antonini, Barausse & Silk 2015a,b). In particular, we discriminate between MBHB mergers driven by gas or by stars by computing the mass of the gas in the central pc-sized nucleus of the galaxy. If the gas mass exceeds the mass of the MBHB, we assume that the merger occurs in a gas-rich environment and we assume that the MBH binary is driven to sub-pc separation (where GW emission is sufficiently efficient to trigger the merger) by planetary-like migration within the nuclear disc. We consider that migration proceeds on the viscous timescale (evaluated at the influence radius of the binary), i.e. $\sim 10^7 - 10^8$ yr. In the opposite case, i.e., when the gas mass is negligible (a situation more relevant for PTAs), the MBH binary is instead driven to sub-pc separations by stellar hardening, i.e., by three-body interactions with stars. We model this phase by assuming that its duration is given by the hardening timescale evaluated at the influence radius of the MBH binary (Sesana & Khan 2015), or to the timescale of the hardening from the nuclear star cluster, whichever the shorter. In practice, these hardening timescales are typically of a few Gyr. We have checked that our results do not depend sensitively on the employed prescription to discriminate gas-rich from gas-poor mergers. In practice, one finds that the vast majority of MBHB mergers is gas-poor in the PTA band (see table 1).

We account for the possibility that while a MBH binary is still evolving under gas-driven migration or hardening, another galaxy merger may take place. This would bring a third MBHs down to pc-separations, which may in turn trigger the coalescence of the inner binary via Kozai-Lidov resonances (Kozai 1962; Lidov 1962) or chaotic three-body

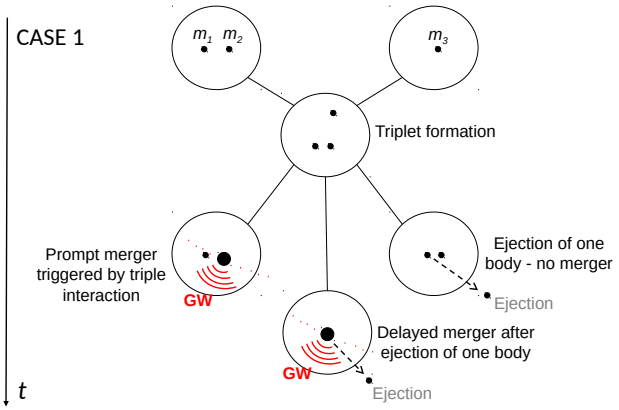

Figure 1. Cartoon representation of how triple MBH interactions are treated in the semianalytic model described in Section 2.

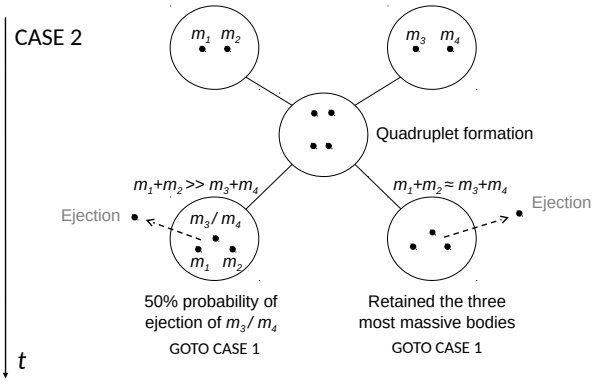

Figure 2. Same as figure 1, but for quadruple interactions.

interactions (Paper I). We model these processes by using the results of the three-body Post-Newtonian (PN) code of Paper I, as outlined in the next section. Clearly, the importance of these MBH triple systems in triggering MBH coalescences depends sensitively on the "delays" between the time halos start merging and the time MBHs eventually coalesce. In the following, we consider both a case where these delays are implemented as described above (*Model-delayed* hereinafter) and a case where the delays are artificially set to values larger than the age of the Universe (i.e., MBHBs stall at about their hardening radius, therefore mergers never take place unless triggered by three-body MBH interactions, *Model-stalled* hereinafter). Coupled with the different MBH seeding prescription, we therefore have four distinct models: *Model-delayed-LS*, *Model-stalled-LS*, *Model-delayed-HS*, *Model-stalled-HS*. We find that low and high mass seeds (*LS* vs *HS*) yield very similar results in the MBH mass range relevant to PTAs (but not at the low masses relevant to LISA, Klein et al. 2016, as we will explore in a forthcoming paper). Unless otherwise stated, we will therefore always show results for the *LS* models.

2.2 Treatment of triple and quadruple MBH systems

The backbone for a consistent treatment of multiple (i.e. triple and quadruple) MBH interactions in the semianalytic model described above is the large suite of numerical sim-

ulations described in Paper II. Full details can be found in there, and we only describe here the key features and the implementation within the semianalytic galaxy evolution model. We have collected the outcome of triple MBH interactions on a grid of primary MBH mass, m_1, sampled in the range $[10^5\mathrm{M}_\odot, 10^{10}\mathrm{M}_\odot]$, and inner and outer binary mass ratio $q_{\rm in} \in [0.03, 1]$and $q_{\rm out} \in [0.03, 10]$[3]. For each point in the 3D grid (m_1, $q_{\rm in}$, $q_{\rm out}$), we simulate several systems with different inner and outer orbit eccentricities, and different relative inclinations, and we use the results to compute merger fractions, merger-time distributions and MBHB eccentricity distributions just before merger (more precisely, we record eccentricities at separations of $100R_{\rm G}$, where $R_G = GM/c^2$ is the gravitational radius associated with the merging MBHB total mass $M = m_1 + m_2$). In particular, we isolate three distinct outcomes (see figure 1) and their associated occurrence probabilities:

(i) A prompt coalescence triggered by a triple interaction. The coalescence can involve any one pair of MBHs in the triplet. We identify here with body 1 and 2 the two MBHs of the inner binary ($m_1 > m_2$ by definition) and with body 3 the intruder. For each grid point in our simulation suite, we record three numbers a, b and c identifying the fractions of simulations in which bodies 1-2, 1-3 and 2-3 merge, respectively.

(ii) An ejection of one of the MBHs (the lighter, in the overwhelming majority of the cases) and a delayed merger of the remaining binary (shrunk or made more eccentric by the 3-body interaction) under the effect of GW emission. This occurs in a fraction d of the realizations.

(iii) An ejection and a left-over binary unable to merge alone within the Hubble time. Such a binary is then retained in the semianalytic model as it can potentially undergo new multiple MBH interactions following later galaxy mergers, or coalesce under the effect of gas-driven migration or stellar hardening (in *Model-delayed* only).

The relative occurrence of the different outcomes depend on the chosen point in the 3-D grid (m_1, $q_{\rm in}$, $q_{\rm out}$). We also stress that our treatment of triple interactions is conservative, because the ejected MBH may fall back to the galactic nucleus after the left-over binary has merged, thus potentially providing an additional MBH merger in a minority ($\sim 10-20\%$) of cases (Hoffman & Loeb 2007).

In any given triple interaction produced by the semianalytic model, the probability of a given outcome is obtained by using a trilinear interpolation between the surveyed grid points, to estimate the fractions a, b, c, d for that specific system. A random number P between 0 and 1 is then drawn and, according to its value, one of the following choices is selected:

- If $P < a + b + c \rightarrow$ prompt merger:
  - If $P \leqslant a \rightarrow$ merger of bodies 1-2.
  - If $a < P \leqslant a + b \rightarrow$ merger of bodies 1-3.
  - If $a + b < P \leqslant a + b + c \rightarrow$ merger of bodies 2-3.
- If $a + b + c < P \leqslant a + b + c + d \rightarrow$ delayed merger (of the two most massive bodies).
- If $a+b+c+d < P \rightarrow$ no merger (left-over binary formed by the two most massive bodies).

In the case of a merger, its timescale is obtained by sampling the distribution of merger timescales for the two cases of prompt and delayed mergers respectively (see Paper II). At this stage, we do not follow self consistently the eccentricity evolution of individual MBHB systems in the semianalytic model (see next section). We also note that the parameters of a given triplet can lie outside the grid sampled in Paper II. In particular we can have $q_{\rm in} < 0.03$ and/or $q_{\rm out} < 0.03$ or $q_{\rm out} > 10$. In this case we simply apply the fractions and timescale distributions of the closest grid point. Although this is certainly a crude approximation, the GW signal is much weaker for low mass-ratio binaries, and our treatment of those systems does not change our results significantly (see section 5).

Besides the formation of triplets, quadruple interactions (caused by the merger of two galaxies, both hosting a binary) are a natural occurrence, especially in *Model-stalled*. In absence of a library of simulations of quadruple interactions, we reduce the problem to the triplet case, as shown in figure 2. If one of the two binaries is much lighter than the other (we arbitrarily choose a threshold mass ratio of 0.1), we expel one of its two members with 50% probability, irrespectively of the binary's mass ratio, retaining only one intruder and reducing the problem to the triplet case. This assumption is made in analogy to the problem of a stellar binary interacting with a much more massive object (usually a MBH or an intermediate MBH, Bromley et al. 2006). If the two binaries have comparable total mass (mass ratio larger than 0.1), we retain the three more massive bodies, again reducing the problem to the triplet case. We stress that this assumption is conservative, mostly because it neglects the possibility of multiple mergers. If the four MBHs form a hierarchical system of two binaries, for example, Kozai-Lidov oscillations might induce mergers of both binaries.

Each of our semianalytic models thus produces a catalogue of MBH mergers containing the masses of the two merging MBHs and the merger's redshift. If the merger involved a standard MBHB, we flag the event either as 'star' or 'gas' depending on whether the binary evolved in a stellar (i.e. gas-poor) or gaseous (i.e. gas-rich) environment. If the merger was instead triggered by a multiple (triple or quadruple) interaction, we also record $q_{\rm in}$ and $q_{\rm out}$ of the progenitor triplet, and we flag the system as 'Tr' if the merger was promptly triggered during the triple interaction, or as 'Tr-ej' if the merger was driven by GW emission after the ejection of one MBH during the triple interaction. In the 'Tr' case, we also record whether the progenitor system was originally a triple or a quadruple, to assess the relative importance of the two populations. Those catalogues are used to construct differential distributions of merging binaries, which we use to compute the GW signal, as detailed in the next section.

[3] While we define $q_{\rm in} \leqslant 1$ by construction, the intruder might be more massive than the pre-existent MBHB (even though this occurs in a minority of cases). Note that for $q_{\rm out} > 1$ we performed a restricted set of simulations for $m_1 = 10^9\mathrm{M}_\odot$ only. Given that none of the results for $q_{\rm out} < 1$ has a strong dependence on the mass scale, we extrapolate the $q_{\rm out} > 1$ results obtained for $m_1 = 10^9$ to all masses.

3 COMPUTATION OF THE GRAVITATIONAL-WAVE SIGNAL

We adopt two different techniques for the computation of the GWB in the case of 'regular' mergers and mergers induced by triple interactions. For regular MBHBs, we assume circular orbits for simplicity. Therefore, the characteristic strain can be expressed as:

$$h_c^2(f) = \frac{4}{\pi f^2} \int dz \int dm_1 \int dq \, \frac{d^3n}{dz dm_1 dq} \frac{1}{(1+z)} \frac{dE_{\rm gw}}{d\ln f_r}. \quad (1)$$

where $d^3n/(dzdm_1dq)$ is the differential number density of MBHB merger per unit redshift, primary MBH mass and binary mass ratio constructed from the output catalogues of the semianalytic model, and the differential energy spectrum $dE_{\rm gw}/d\ln f_r$ can be broken down as

$$\frac{dE_{\rm gw}}{d\ln f_r} = \frac{dE_{\rm gw}}{dt_r}\frac{dt_r}{df_r} f_r. \quad (2)$$

Note that time and frequency are evaluated in the source rest-frame, so that compared to the time and frequency at the observer, we have $t_r = t/(1+z)$ and $f_r = f(1+z)$. The first term on the right-hand side of equation (2) is the GW luminosity, given by

$$\frac{dE_{\rm gw}}{dt_r} = \frac{32}{5}\mathcal{M}^{10/3}(\pi f_r)^{10/3}, \quad (3)$$

where $\mathcal{M} = (m_1m_2)^{3/5}/(m_1+m_2)^{1/5}$ is the binary's chirp mass. The term dt_r/df_r represents the time a given binary spends emitting at a specific frequency f_r. The main contributors to the GWB are moderately heavy binaries merging in massive galaxies at relatively low redshift. Since most of those mergers are gas poor (cf table 1), we assume that MBHBs evolve exclusively because of three-body interactions against the stellar environment and emission of GW, and we can write

$$\frac{df_r}{dt_r} = \frac{df_r}{dt_r}\Big|_{3b} + \frac{df_r}{dt_r}\Big|_{\rm gw} = \mathcal{A}f^{1/3} + \mathcal{B}f^{11/3}, \quad (4)$$

where (Chen, Sesana & Del Pozzo 2017)

$$\mathcal{A} = \frac{3}{2\pi^{2/3}}\frac{H\rho_i}{\sigma}M^{1/3}, \quad (5)$$

$$\mathcal{B} = \frac{96}{5}(\pi)^{8/3}\mathcal{M}^{5/3}. \quad (6)$$

Here, $M = m_1 + m_2$ is the MBHB total mass, σ is the velocity dispersion of the host galaxy, ρ_i is the stellar density at the binary's influence radius (both of which are evaluated from the output of the semianalytic galaxy-formation model), and $H \approx 15$ is a numerical factor calibrated against dedicated three-body scattering experiments (Quinlan 1996; Sesana, Haardt & Madau 2006). The binary evolution is dominated by three-body scattering in the early phase, whereas GW emission takes over at higher frequencies. The transition frequency f_t can be computed by equating the two contributions to df/dt, and for typical PTA sources lies in the nHz range.

Triple interactions can result in extremely eccentric binaries, making the analytic computation of the GWB somewhat time consuming (see however, Taylor, Simon & Sampson 2017). Moreover, the eccentricity evolution can be highly chaotic (Paper I), and thus the construction of a simple analytic dE/df is not possible. We note, however, that in the PTA frequency range MBHBs show a rather regular behaviour, following $f-e$ tracks dictated by GW backreaction and hardly affected by the third body (cf figure 9 in Paper II). We can therefore consider, to first approximation, a population of eccentric MBHBs evolving solely because of GW emission. Chen, Sesana & Del Pozzo (2017) showed that the stochastic GWB for an arbitrary population of GW-driven eccentric binaries can be simply obtained by evaluating the single spectrum of a fiducial system, and rescaling it appropriately to match the parameter of the considered sources. The total GWB can be written as:

$$h_c^2(f) = \int dz \int dm_1 \int dq \int de \frac{d^4n}{dz dm_1 dq de} h^2_{c,\rm fit}\Big(f\frac{f_{p,0}}{f_p}\Big) \Big(\frac{f_p}{f_{p,0}}\Big)^{-4/3}\Big(\frac{\mathcal{M}}{\mathcal{M}_0}\Big)^{5/3}\Big(\frac{1+z}{1+z_0}\Big)^{-1/3}, \quad (7)$$

where $d^4n/(dzdm_1dqde)$ is now the differential number density of MBHB mergers per unit redshift, primary mass, mass ratio and eccentricity. This quantity is constructed from the catalogue produced by the semianalytic model. Since we do not follow the eccentricity evolution of MBHBs self-consistently, for each event we draw a value of e from the distribution corresponding to the appropriate parent triplet properties, interpolating the distributions obtained at the grid points of our suite of numerical integrations (cf section 2.2; as mentioned earlier, eccentricity distributions are recorded at a reference binary separation of $100R_{\rm G}$). $h_{c,\rm fit}$ is a reference spectrum for a binary with parameters $(\mathcal{M}_0, z_0, f_0, e(f_0))$, which is adapted to arbitrary MBHB parameters via the scaling factors reported in parenthesis. The factor $f_{p,0}/f_p$ is the ratio of the peak frequencies of the two binary spectra. An eccentric binary, in fact, has a peak in the emission spectrum that is uniquely determined by specifying $e(f)$ at a given frequency f (details given in Chen, Sesana & Del Pozzo 2017). Therefore, if we know e at $100R_G$, we can compute f_p and rescale the fiducial spectrum accordingly. We also recall that triple interactions can result in either a 'Tr' or 'Tr-ej' merger, as described in the previous section. In general, 'Tr' and 'Tr-ej' events have very different eccentricity probability distributions, and we therefore distinguish between the two cases and sample from the respective distributions.

Although qualitatively different, the two GWB computations of equations (1) and (7) are perfectly consistent with each other. We have checked that by artificially setting arbitrarily small e in the triplet population, the GWB obtained via equation (7) coincides with that obtained via equation (1), assuming purely GW-driven circular binaries.

Summarising, to practically evaluate the GWB in the two models, we proceed as follow. We flag the origin of each MBHB as merging because of:

(i) Standard dynamical processes (flag 'star').
(ii) Dynamical processes during a triple interaction (flag 'Tr').

(iii) GW emission after the ejection of the lighter MBH involved in the triple interaction (flag 'Tr-ej').

For each subset of systems we construct the relevant differential number density $d^3n/(dzdm_1dq)$ for case (i), or $d^4n/(dzdm_1dqde)$ for cases (ii) and (iii). In *Model-delayed*, all subsets contribute to the GWB, and we therefore write $h_c^2 = h_{c,\rm star}^2 + h_{c,\rm Tr}^2 + h_{c,\rm Tr-ej}^2$, where $h_{c,\rm star}^2$ is computed via equation (1) and $h_{c,\rm Tr}^2$, $h_{c,\rm Tr-ej}^2$ are obtained via equation (7). In *Model-stalled*, only triple interactions can drive MBHB coalescences; the GWB is therefore computed as $h_c^2 = h_{c,\rm Tr}^2 + h_{c,\rm Tr-ej}^2$, where both terms are obtained via equation (7).

4 RESULTS

4.1 MBH merger rates

We first focus on the MBHB merger rates predicted by *Model-delayed* and *Model-stalled*. Being interested primarily in the PTA signal, we concentrate on systems with $\mathcal{M} > 10^7\rm M_\odot$. Results are reported for both *HS* and *LS* models in table 1 and plotted for the *LS* model (our default choice) in figure 3. The table lists the total merger rates for selected chirp-mass ranges, and highlights the relative contributions of different MBHB sub-populations, providing a number of useful information:

(i) As anticipated, there is little difference between the *HS* and *LS* models; rates are very similar for all mass ranges and so are the fractions of mergers due to individual channels.

(ii) The only exception is the fraction of mergers due to quadruple interactions, which is larger in the *LS* model. This is due to this model's large occupation fraction of MBHs with $M < 10^5\rm M_\odot$, which is also responsible for the presence of very low mass ratio binaries (DB17), which are absent in the *HS* model. We stress, however, that these low mass ratio systems have little effect on the level of the GWB.

(iii) In *Model-delayed*, more than 75% of all merging systems with $\mathcal{M} < 10^9\rm M_\odot$ are 'regular' binaries, the vast majority of which reside in gas poor environments. This validates our assumption that regular binaries evolve via stellar hardening only.

(iv) The importance of triple interactions is a monotonically increasing function of mass. For $\mathcal{M} > 10^9\rm M_\odot$, about half of the mergers are due to this channel. This is because very massive galaxies experience several mergers in their lifetime, hence typical MBHB merger timescales are longer than the time occurring between subsequent galaxy mergers. *Model-delayed* is therefore similar to *Model-stalled* at such high masses, but we checked that systems with $\mathcal{M} > 10^9\rm M_\odot$ (which are quite rare, see the 'Rate' column in the table) contribute less than 10% to the overall GWB signal.

(v) For $\mathcal{M} < 10^9\rm M_\odot$, merger rates of *Model-delayed* are about four times higher than those of *Model-stalled*. Triple interactions have therefore limited efficiency (about 30%) in solving the final-parsec problem, since many of them simply end up with the ejection of one of the MBHs (about 70%) without the left-over binary merging. Note that, conversely, rates are comparable for $\mathcal{M} > 10^9\rm M_\odot$, for the reasons discussed in the previous point.

(vi) In general, $\approx 80-90\%$ of triplet-induced mergers are prompt (Tr), whereas only about $10-20\%$ are due to GW hardening following the ejection of the lighter of the three MBHs (Tr-ej). This is true for both *Model-delayed* and *Model-stalled*.

The redshift distribution of merging systems is shown in figure 3 for the LS model. Counter intuitively, mergers do not appear to be shifted, on average, to lower redshifts in *Model-stalled*. This is likely because typical MBHB merger timescales in *Model-delayed* are of the order of Gyrs, especially when systems are stellar driven (which are the vast majority at $\mathcal{M} > 10^7\rm M_\odot$), similar to the timescale of subsequent mergers that trigger triple interaction. Within the triple-induced mergers, however, the Tr-ej sub-group (dotted curves in figure) tends to coalesce at lower redshifts than the Tr one (dashed curves in the figure). This is because the former is comprised of left-over systems that merge because of GW emission only, and their coalescence timescale is skewed towards values of several Gyrs (cf Paper II, figure 7), thus shifting the peak of the merger rate to lower z. The Tr binaries, conversely, typically coalesce in few hundred Myrs. The overall shape and normalisation of the rates are in line with estimates from other authors (e.g., Blecha et al. 2016), and the implied total merger rate of MBHBs with $\mathcal{M} > 10^7\rm M_\odot$ is about 0.14 $\rm yr^{-1}$ in *Model-delayed* and 0.032 $\rm yr^{-1}$ in *Model-stalled*.

4.2 Stochastic GW background

Figure 4 shows the stochastic GWB produced by the two models. The figure is obtained by combining 100 Monte Carlo sampling of the $d^4n/(dzdm_1dqde)$ distribution. To assess the overall effect of MBHB stalling in the normalisation of the expected GWB, we first ignore effects due to stellar hardening and eccentricity and compute the GWB as due to circular GW-driven binaries. In this case, standard calculations gives $h_c = A[f/(1\ \rm yr^{-1})]^{-2/3}$, shown as a golden line in figure 4. In practice, we evaluate the integrals in eq. 1 under the assumption that df_r/dt_r is solely determined by the (circular) emission of GWs (i.e., we set $\mathcal{A} = 0$ in eq. 4). We find $A = 1\times10^{-15}$ and $A = 0.7\times10^{-15}$ for *Model-delayed* and *Model-stalled* respectively.

In *Model-delayed*, the vast majority of the GWB ($h_{c,\rm star} \approx 0.9h_c$) is produced by regular binaries evolving via stellar hardening, with triplet-induced mergers (either prompt or following ejection) playing a sub-dominant role. As long as MBHBs are not highly eccentric, the spectral turnover is at $f < 1$ nHz (black and green lines in the left panel of figure 4), and the signal only mildly departs from the $f^{-2/3}$ power law at frequencies relevant to PTA detection. This is certainly true in our model, which assumes circular binaries, but the shape of the spectrum is hardly affected by eccentricity up to $e \approx 0.5$ (Sesana 2015; Taylor, Simon & Sampson 2017; Chen, Sesana & Del Pozzo 2017). We note, however, that the evolution of the eccentricity of stellar driven binaries strongly depends on its initial value at binary formation, which is a poorly understood parameter, with N-body simulations of merging galactic bulges resulting in a wide range of MBHB eccentricities (see Dotti, Sesana & Decarli 2012, for a review). Conversely, the signal in *Model-stalled* already departs from the $f^{-2/3}$ power law at $f \approx 3\times10^{-8}$ Hz, and at 1 nHz it is already a factor ~ 2 below

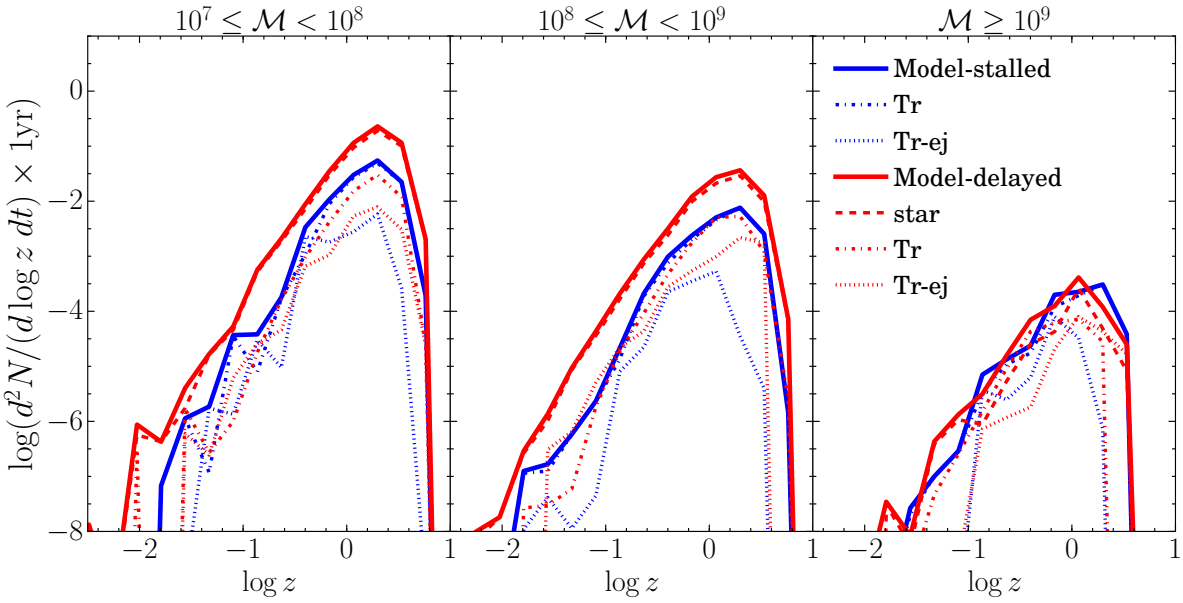

Figure 3. Redshift distribution of merging MBHBs in different chirp mass bins, as labelled at the top of the right panel. Line styles are described in the figure.

LS										
$\mathcal{M}$ [$M_\odot$]	*Model-delayed*					*Model-stalled*				
	Rate [yr^{-1}]	binaries		triplets		Rate [yr^{-1}]	binaries		triplets	
		star	gas	Tr (quad)	Tr-ej		star	gas	Tr (quad)	Tr-ej
$10^7 - 10^8$	0.118	80.0%	5.1%	12.4%(1.3%)	3.6%	0.028	–	–	89.3%(77.3%)	10.7%
$10^8 - 10^9$	0.022	75.9%	2.9%	15.0%(3.4%)	6.2%	4.4×10^{-3}	–	–	93.6%(81.8%)	6.3%
10^9	1.8×10^{-4}	50.7%	0.04%	28.0%(11.2%)	21.3%	1.9×10^{-4}	–	–	84.6%(77.4%)	15.4%
HS										
$\mathcal{M}$ [$M_\odot$]	*Model-delayed*					*Model-stalled*				
	Rate [yr^{-1}]	binaries		triplets		Rate [yr^{-1}]	binaries		triplets	
		star	gas	Tr (quad)	Tr-ej		star	gas	Tr (quad)	Tr-ej
$10^7 - 10^8$	0.079	79.2%	4.2%	13.0%(2.2%)	3.6%	0.044	–	–	89.3%(64.7%)	10.7%
$10^8 - 10^9$	0.020	76.7%	4.3%	15.1%(3.5%)	3.9%	5.4×10^{-3}	–	–	86.6%(65.8%)	13.4%
10^9	2.4×10^{-4}	72.2%	2.1%	21.3%(4.8%)	4.5%	1.9×10^{-4}	–	–	79.9%(65.6%)	20.1%

Table 1. Merger rate and population composition of the MBHBs with chirp mass in the three most massive mass bins. The number in parenthesis refer to the fraction of prompt mergers originated by a quadruple system.

its nominal $f^{-2/3}$ value. This is due to the non-negligible eccentricity of MBHBs merging via triple interactions. Unlike in the stellar hardening scenario, the presence of binaries with high eccentricities is not just a possibility in this case, but an *inevitable* outcome of the three-body MBH dynamics (cf Paper I and Paper II). Note the relative contribution of promptly induced coalescences ($h_{c,\mathrm{Tr}}$) and GW-driven coalescences following ejection of one of the triplet's members ($h_{c,\mathrm{Tr-ej}}$). The normalisation of the latter contribution is a factor ~ 2 lower, being Tr-ej systems about 20% of the overall triplet-induced coalescences. However, the contributions of the two sub-populations have different spectral shapes, crossing at $f \approx 3 \times 10^{-9}$Hz, below which Tr-ej becomes dominant. This is due to the different eccentricity distribution of the two sub-populations, as we will see in the next subsection.

Figure 5 visualises the difference between the simple $f^{-2/3}$ power law, *Model-delayed* and *Model-stalled*. Results are shown for both *LS* and *HS* models, to stress their similarity. The plot clearly shows that *Model-delayed* closely follows the simple power law model at least down to 2×10^{-9}Hz, with a low-frequency drop due to stellar driven evolution. The ratio between *Model-stalled* and the single power-law model, as already mentioned, is about 0.7 at high frequencies, monotonically decreasing to about 0.1 at 0.1 nHz. Compared to *Model-delayed*, *Model-stalled* produces a GWB that is a fraction 2-to-3 smaller in the frequency range 1-10 nHz, most relevant to PTA experiments. The result holds for both *LS* and *HS* models, with minimal differences.

4.2.1 Eccentricity distribution

As mentioned above, the occurrence of high eccentricities is a critical and inevitable feature of triple-induced inspirals. This might be relevant for the detection of individually resolvable binaries, for which eccentric templates might be

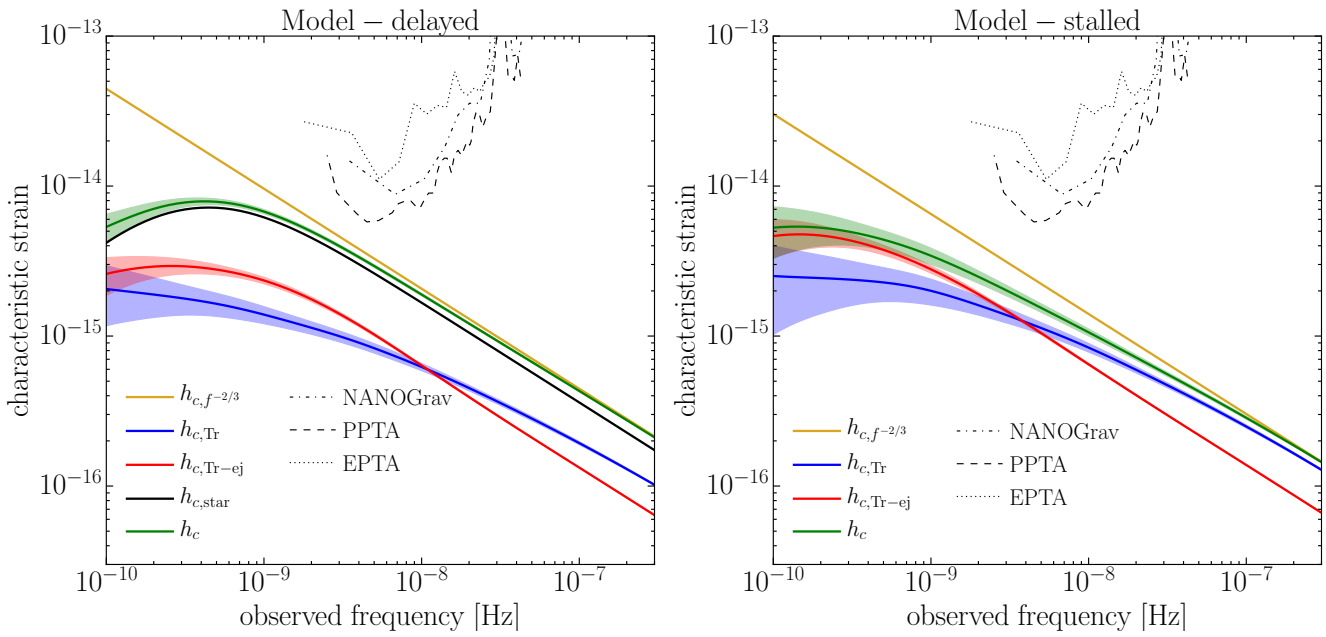

Figure 4. Stochastic GWB from MBHBs, divided in each of the contributing components, as indicated in each panel. Solid lines show the mean values of the characteristic strain h_c of each component, whereas the shaded area marks the standard deviation inferred from 100 Monte Carlo simulations of the MBHB population. Current sensitivities of major PTA experiments are also shown.

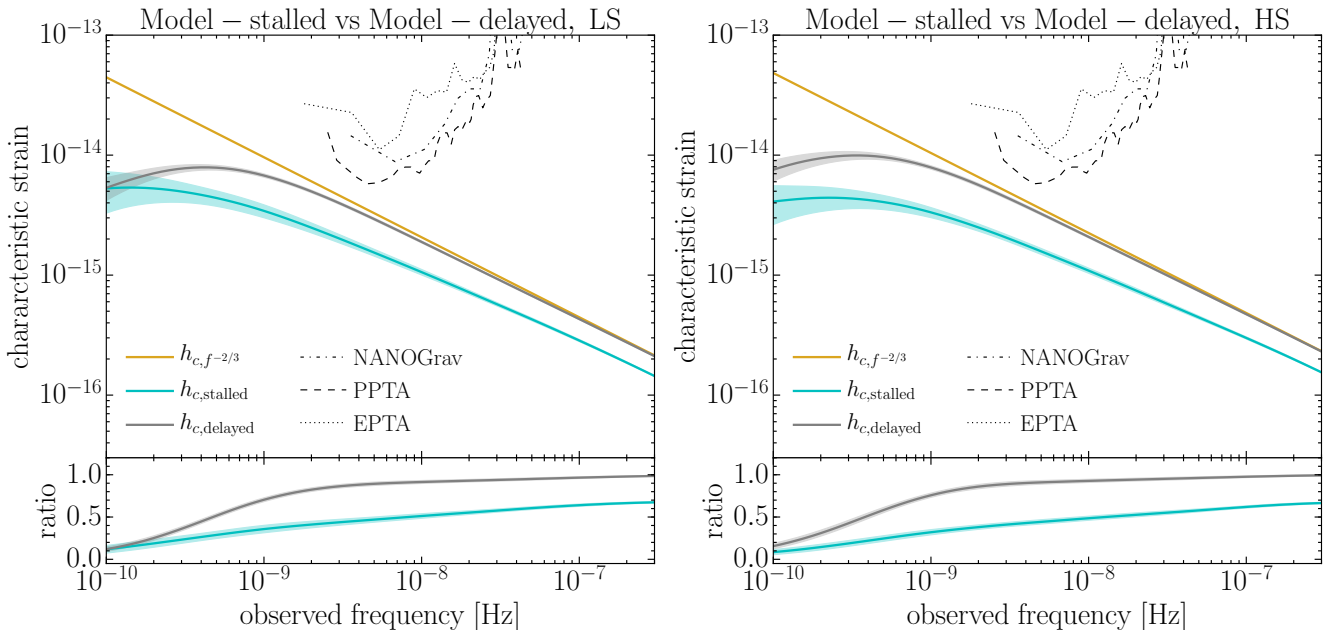

Figure 5. Comparison of the stochastic GWB generated by *Model-delayed* (grey) and *Model-stalled* (light-blue). Lines and shaded areas have the same meaning as in figure 4. The bottom inset shows the ratio between either models and a reference $f^{-2/3}$ power law generated by a GW-driven population of circular binaries, shown as an orange line in the main plots. Here, we present results for both LS (left) and HS (right) seed models.

necessary (Taylor et al. 2016a). To investigate the actual distribution of individual MBHB eccentricities, we need to convert the merger number densities in equation (7) into the instantaneous number of systems in the sky at a given frequency, i.e.

$$\frac{d^5N}{dzdm_1dqded\ln f} = \frac{d^4n}{dzdm_1dqde'}\frac{dV_c}{dz}\frac{dz}{dt}\frac{dt}{d\ln f}, \qquad (8)$$

where e' is the eccentricity computed at $100R_G$, V_c is the comoving volume, dV_c/dz, dz/dt are known once a cosmology

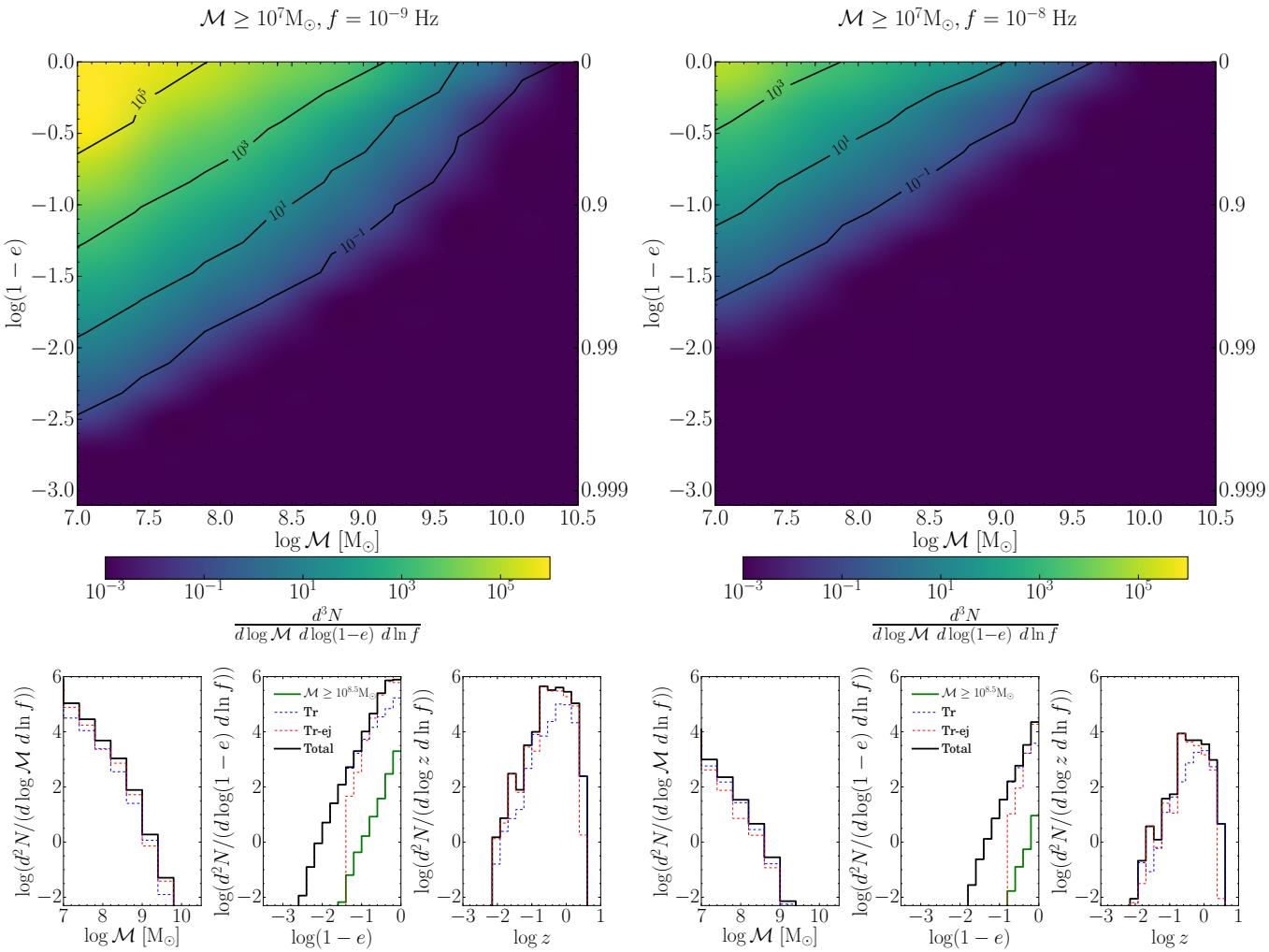

Figure 6. Properties of individual MBHBs contributing to the GW signal within an observed orbital frequency $\Delta f = f$ around $f = 1$ nHz (left) and $f = 10$ nHz (right). The upper panels show the differential distribution of sources in the chirp mass – circularity plane (with the eccentricity e shown on the right of each figure). The lower panels show the marginalised distributions of the number of sources as a function of chirp mass (left), circularity (centre) and redshift (right). The legend of the histogram linestyles is shown in the central panels.

is assumed (Hogg 1999), and

$$\frac{dt}{d\ln f} = \frac{5}{96}(2\pi)^{-8/3}\mathcal{M}^{-5/3}f^{-8/3}F(e)^{-1} \quad (9)$$

$$F(e) = \frac{1 + (73/24)e^2 + (37/96)e^4}{(1-e^2)^{7/2}}. \quad (10)$$

The eccentricity e at the desired frequency f is obtained by evolving e' backwards from $100R_G$ to f by using the standard evolution of eccentric binaries in the quadrupole approximation (Peters & Mathews 1963; Chen, Sesana & Del Pozzo 2017). Note that $F(e)$ is a strong increasing function of e, thus $dt/d\ln f$ is much shorter for very eccentric binaries, down-weighing their relative number at a given observed frequency.

The distribution of the number of emitting binaries, integrated in redshift, in the circularity[4]-mass plane is shown in figure 6 at two observed *orbital* reference frequencies $f = 1$nHz and $f = 10$nHz. The number of sources is obviously dominated by low mass MBHBs, with a long tail of few sources extending up to $\mathcal{M} = 10^{10}\mathrm{M}_\odot$. The overall distribution is dominated by light, rather circular binaries, but possible eccentricities extend up to $e > 0.99$.

Marginalised source distributions are shown in the lower panel as a function of chirp mass, circularity and redshift. As expected, the number of sources is dominated by low-mass systems (which, however, do not contribute much to the total GWB) and the redshift distribution peaks around $z \sim 1$, consistent with, e.g, Sesana, Vecchio & Volonteri (2009) and Blecha et al. (2016). As already noticed, the Tr and Tr-ej sub-populations behave quite differently, the latter peaking at lower redshifts (because of the long coalescence timescales). Moreover, the circularity distributions are also distinct: the Tr population extends to $1 - e \approx 0.003(0.03)$ at $f = 1(10)$ nHz, whereas the the Tr-ej populations hardly goes below $1 - e \approx 0.1$, preferentially selecting rather circular binaries. This can be understood by looking at figure

[4] The circularity is defined as 1-e. Its logarithm is often used for plotting purposes, to highlight tails of high eccentricities.

10 of Paper II. Prompt coalescences (i.e. the Tr population) are generally caused by the formation of either a temporarily highly eccentric binary (mostly as a result of secular Kozai-Lidov oscillations), or a compact system of moderate eccentricity (in the case of chaotic energy and angular momentum exchanges). The resulting eccentricity distribution is therefore extremely broad, spanning more than six orders of magnitude at the innermost circular stable orbit. Conversely, coalescences driven by GW emission after an ejection (i.e. the Tr-ej population) preferentially select the systems that did not reach sufficiently high eccentricities to promptly coalesce, but which are still sufficiently eccentric that their coalescence time is shorter than the Hubble time. The result is a much narrower (and mass dependent) allowed eccentricity range, which does not reach the high values of the Tr population. The green curves in the two central panels select binaries with $\mathcal{M} > 10^{8.5}\mathrm{M}_{\odot}$, which are the loudest GW sources and which are thus more likely to be individually resolved. Although the distribution favours circular binaries, $\mathcal{O}(10)$ systems have eccentricities higher than 0.7 at $f \approx 1$ nHz. Eccentric resolvable sources are therefore a rather common occurrence if merging MBHBs are mostly driven by triple interactions.

4.3 A realistic lower bound to the GWB: implications for PTA detectability

Our results imply that triple interactions efficiently counteract the effect of stalling, driving MBHBs to coalescence and resulting in a GWB being reduced by a factor of 2-to-3 only in the PTA band. We can now use this fact to derive a robust lower limit to the amplitude distribution of the expected GWB, based on our best astrophysical knowledge of MBH assembly and dynamics following galaxy mergers.

Our scope is not to make realistic predictions of time to detection (see, e.g. Siemens et al. 2013; Taylor et al. 2016b), but to assess the impact of the most pessimistic scenario on GWB detectability by PTAs. As such, we consider a simple model based on the following assumptions:

(i) We take a GWB amplitude distribution predicted by a MBHB population model based on the MBH-galaxy scaling relations proposed by Shankar et al. (2016), which are based on a putative observational selection bias on the resolvability of the MBH influence sphere. This choice is solely based on the fact that those relations provide a very conservative estimate of the stochastic GWB. In fact, Sesana et al. (2016) showed that they result in a 95% confidence GWB amplitude distribution in the range $1.4\times10^{-16} < A < 1.1\times10^{-15}$, with a median value of $A = 4\times10^{-16}$, well below current PTA limits. Also, consistently with this choice, the semianalytic galaxy-formation model utilised in this paper reproduces the scaling relations of Shankar et al. (2016) (cf. Barausse et al. 2017).

(ii) We draw A from this distribution and apply a correction factor $C(f)_{\rm delayed} = h_{c,\rm delayed}(f)/h_{c,f^{2/3}}(f)$ and $C(f)_{\rm stalled} = h_{c,\rm stalled}(f)/h_{c,f^{2/3}}(f)$, shown in the lower panels of figure 5. In both cases, the GWB then takes the form $h_{c,X} = C(f)_X \times A(f/1\ \mathrm{yr}^{-1})^{-2/3}$.

(iii) We consider an idealised SKA-type PTA. Following Janssen et al. (2015), we make the conservative assumption that SKA will be able to monitor up to 50 MSPs with rms precision of 100ns. We then explore the detection probability (DP) as a function of observing time T and number of pulsars N_p in the array.

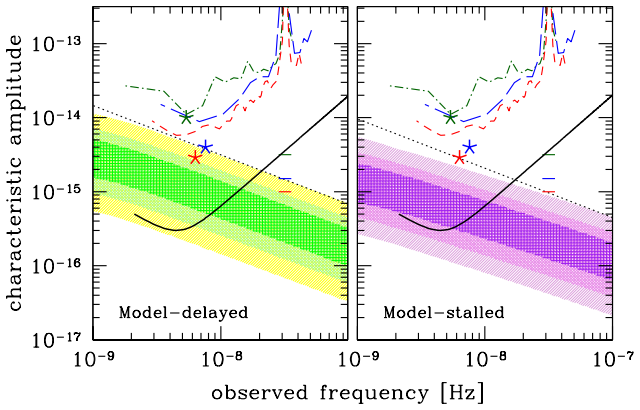

Figure 7. h_c vs f for *Model-delayed* (left panel) and *Model-stalled* (right panel) assuming the conservative MBH population model from Shankar et al. (2016). In each panel, the shaded areas represent the 68%, 95% and 99.7% confidence intervals of the signal amplitude. The jagged curves are current PTA sensitivities: EPTA (dot-dashed green), NANOGrav (long-dashed blue), and PPTA (short-dashed red). For each sensitivity curve, stars represent the integrated upper limits to an $f^{-2/3}$ background, and the horizontal ticks are their extrapolation at $f = 1\ \mathrm{yr}^{-1}$. The solid black line represent the typical sensitivity level of a conservative SKA-type array formed by 50 pulsars monitored at 100ns precision for 15 years. A dotted black line with slope $f^{-2/3}$ is also added to guide the eye.

In particular, assumptions (i) and (ii) provide a realistic projection of how low the GWB can get, by combining the existence of moderately light MBHBs to stalling.

PTA detectability under assumption (iii) is computed using the framework developed by Rosado, Sesana & Gair (2015). If the noise is described by a Gaussian distribution with zero mean and variance σ_0, the DP of a stochastic signal is also described by a Gaussian distribution with mean μ_1 and dispersion σ_1 given by

$$\gamma_{\rm B} = \frac{1}{2}\mathrm{erfc}\left(\frac{\sqrt{2}\sigma_0\mathrm{erfc}^{-1}(2\alpha_0) - \mu_1}{\sqrt{2}\sigma_1}\right), \qquad (11)$$

where erfc is the complementary error function and α_0 is the chosen false alarm probability threshold, which we fix at 0.001. Under our simplifying assumptions of an array of equal pulsars, randomly distributed in the sky and monitored for the same timespan T, μ_1 σ_0 and σ_1 take the form

$$\mu_1 = N_p(N_p-1)T\int df\,\frac{\Gamma^2 S_h^2}{(P+S_h)^2+\Gamma^2 S_h^2}, \qquad (12)$$

$$\sigma_0^2 = N_p(N_p-1)T\int df\,\frac{\Gamma^2 S_h^2 P^2}{[(P+S_h)^2+\Gamma^2 S_h^2]^2}, \qquad (13)$$

$$\sigma_1^2 = N_p(N_p-1)T\int df\,\frac{\Gamma^2 S_h^2\left[(P+S_h)^2+\Gamma^2 S_h^2\right]}{[(P+S_h)^2+\Gamma^2 S_h^2]^2}, \qquad (14)$$

where we have made the further assumption that the signal's spectrum S_h is known and matched to a template $S_{h0} = S_h$, and we have replaced the pulsar-pair dependent correlation function Γ_{ij} (also known as Hellings & Downs function,

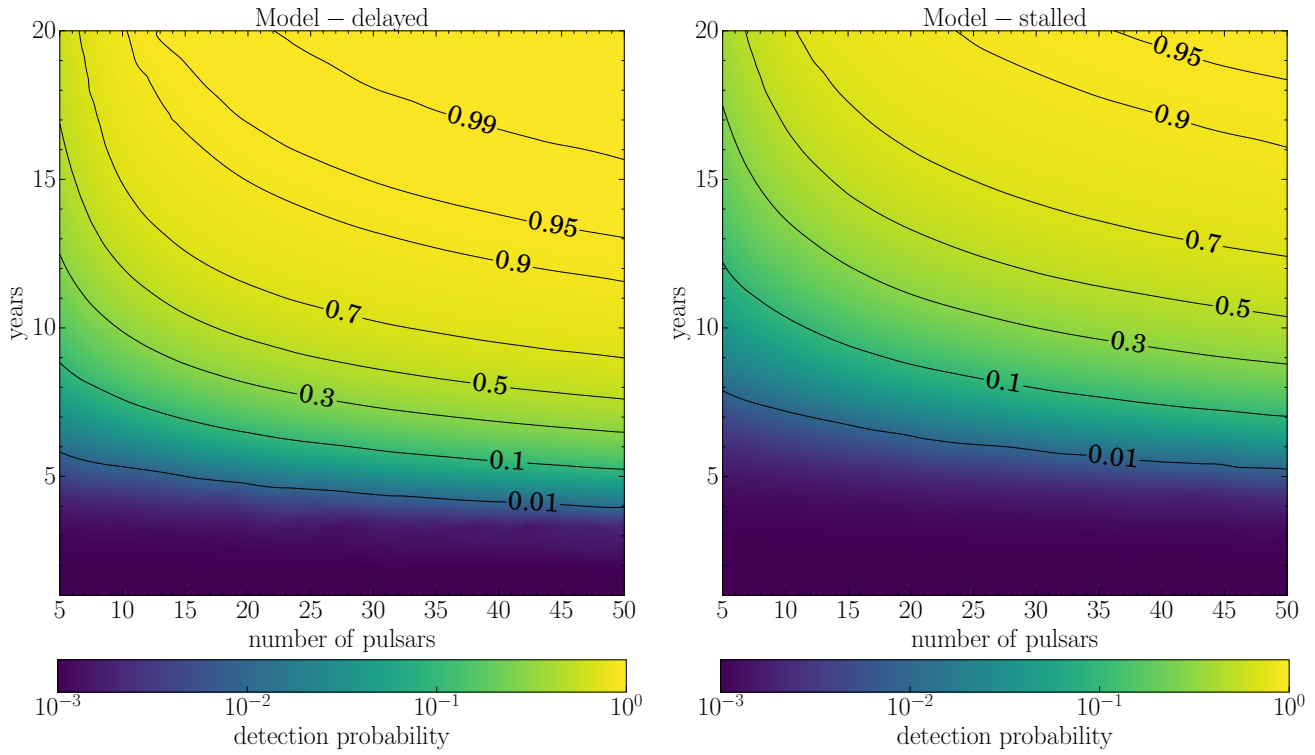

Figure 8. Detection probability contour plot in the number of pulsar vs observation time plane. All pulsars are assumed to have an equal rms residual of 100ns.

Hellings & Downs 1983) with the square root of its variance, i.e., $\Gamma = 1/(4\sqrt{3})$. The signal's power spectral density (PSD) S_h is related to the characteristic strain derived in Section 3 via:

$$S_h = \frac{h_c^2}{12\pi^2 f^3}. \tag{15}$$

For the PSD of the noise P, assumed to be the same for all pulsars, we use the form

$$P = 2\sigma^2 \Delta t + \frac{\delta}{f^5}, \tag{16}$$

where $\sigma = 100$ns is the rms residual of the measured TOAs, Δt is the assumed cadence of individual MSPs observations, and

$$\delta = 5 \times 10^{-49} \left(\frac{10\text{yr}}{T}\right)^5 \left(\frac{\sigma}{100\text{ns}}\right)^2 \frac{\Delta t}{2\,\text{weeks}}. \tag{17}$$

With this prescription, the second term on the right-hand side of equation (16) mimics the loss of sensitivity seen in current PTAs at low frequency and due to fitting the MSP spin and spin derivative in the timing model. Note that we do not include any red-noise contribution to the noise PSD, which can be easily accounted for by adding a suitable term $P_{\rm rn}(f)$ in equation (16). For each value of A drawn from the GWB-amplitude distribution reported in Sesana et al. (2016), we compute the expected $h_c(f)$ for *Model-delayed* and *Model-stalled* as explained in point (ii) above, and for each value of N_p and T we compute the DP using equation (11).

Results are shown in figures 7 and 8. Figure 7 shows the distribution of the expected amplitudes as a function of frequency. In *Model-stalled*, the signal is generally flatter than the canonical $f^{-2/3}$ power law in the relevant PTA frequency range, thus it is not sufficient to simply report GWB amplitudes A when quoting results. This is also true in *Model-delayed*, even though the departure from $f^{-2/3}$ is minimal at $f > 1$nHz. The *Model-stalled* amplitude range still spans more than an order of magnitude and is shifted down by about a factor of two at the currently most relevant PTA frequencies (marked by the stars) compared to the fiducial model. The whole predicted range (99.7% confidence region) is below the current best PTA limit (Shannon et al. 2015), but well within the reach of a putative SKA array under our conservative assumptions.

This is better quantified in figure 8, that shows the DP in the observation time (T) – number of pulsars (N_p) plane. For a given N_p, the 50% DP timescale is delayed by only 3-to-6 years in the *Model-stalled* scenario. The plot highlights the importance of having a sufficiently large N_p, i.e. a larger array helps to narrow this time gap. In fact, detection is based on correlation statistics, which is very sensitive to the number of pulsar pairs that can be correlated. We see that if $N_p = 5$, even at $T = 20$ yr we still have only DP ≈ 0.3. A larger array of $N_p = 50$, instead, reaches the same DP value after only 10yr and by $T = 20$ yr has DP > 0.95. Overall, these findings support the statement that PTAs *will* eventually detect the stochastic GWB from MBHBs, regardless of possible binary stalling issues.

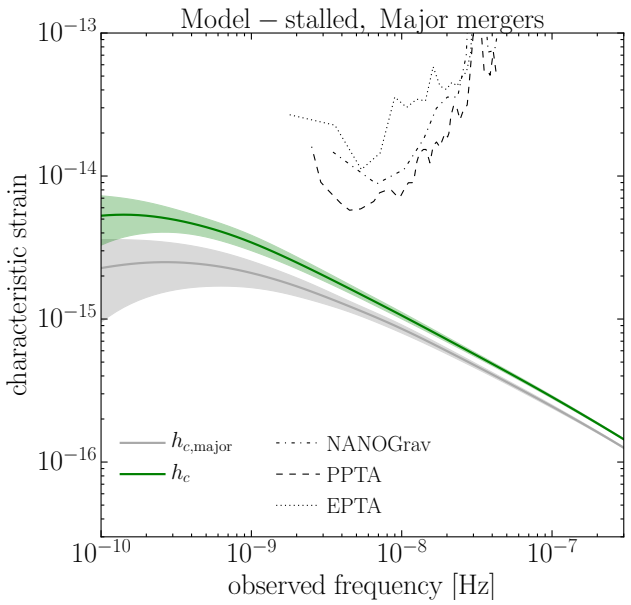

Figure 9. GWB spectrum h_c of *Model-stalled* when considering only major mergers, i.e., when $q_{\rm in} \geqslant 0.1$ and $q_{\rm out} \geqslant 0.1$ (grey line and area) compared to the total GWB predicted by the model (green line and area). Lines and shaded areas have the same meaning as in figure 4.

5 CAVEATS

Although employing an accurate treatment of the three-body dynamics including an external potential, dynamical friction, stellar hardening and PN equations of motion consistently derived from the 3-body PN Hamiltonian (Paper I), the results presented here are subject to a number of caveats that we briefly discuss in the following.

First, we did not attempt to model the interaction of quadruple MBH systems. This is an important point, especially in *Model-stalled* where, necessarily, most galaxy mergers contain pairs of MBHBs instead of pairs of single MBHs. In fact, in this case mergers are dominated by quadruple systems (cf table 1). As already mentioned, by removing one body and thus recovering a triplet, our estimation of the quadruple contribution to the GWB is conservative. We also stress that even if for some (unexpected) reasons quadruple interactions were to lead to no mergers, the total GWB would be further suppressed by a factor of two only, and we therefore conclude that our results are robust against this instance.

Second, in the simulations of Paper II we did not account for the later evolution of the MBHs that are ejected in triple systems. As shown by Hoffman & Loeb (2007), those MBHs may fall back to the galactic nucleus on timescales shorter than the Hubble time, thus giving rise to an additional MBH merger in about $10-20\%$ of cases. For this reason, however, this effect is sub-dominant relative to the main one (the prompt and GW-driven triplet-induced mergers that we account for in this paper), and is likely to depend on the exact modelling of the galactic potential (and namely its triaxiality, which is unknown). Moreover, as already mentioned, since we neglect this effect our results should be regarded as conservative.

Third, as stressed in section 2.1, our semianalytic galaxy-formation model only includes the dynamical friction timescale of galaxy satellites in the potential well of the primary galaxy until the two merge, but does not model the the early migration of MBHs driven by dynamical friction against the gaseous and stellar distribution, in the early epochs following the merger. We do this on the grounds that this timescale is generally short relative to that of the dynamical friction between the two halos and galaxies, and relative to the timescales that describe the evolution of MBHBs at separations $\lesssim$ a few pc (stellar hardening, gas-driven migration, triple MBH interactions), at least for the comparable-mass MBHBs that provide the bulk of the PTA signal (DB17, Dosopoulou & Antonini 2017). For this reason, the results of this paper are robust against this assumption, which anyway affects only the *Model-delayed* results and not directly the *Model-stalled* ones (where the MBH merger timescales are set to values larger than the Hubble time). Nevertheless, N-body simulations of galaxy pairs find that merging within an Hubble time might be difficult for MBH systems with mass ratios $\lesssim 0.1$. Therefore this issue, while formally absent in *Model-stalled*, may have consequences also for that scenario, because the simulations of Paper II *assume* (as initial conditions) that MBHs are efficiently brought down to separations comparable to the primary MBH's sphere of influence. However, we have checked that low mass ratio systems do not contribute significantly to our results, even in *Model-stalled*, by removing all triplets with either $q_{\rm in}$ or $q_{\rm out}$ lower than 0.1 from the GWB calculation. Results are shown in figure 9. It is clear that low q systems do not significantly contribute to the GWB normalisation, being the signal at high frequencies only about 10% lower after their removal. It appears, however, that high q triplets tend to produce more eccentric binaries, causing a higher frequency flattening and spectral turnover when low q systems are not included in the calculation. The difference in GWB amplitude is however still less than 50% at frequencies of few nHz, relevant to PTA experiments.

Last, when applying our results to the Sesana et al. (2016) GWB signal distribution, we are implicitly assuming that the signal correction factor $C(f)$ due to stalling directly applies to MBHB populations that are different from those produced by our semianalytic model. Indeed, even though the two models both assume observational selection effects on the scaling relations, the intrinsic MBH-galaxy scaling relations of the populations are not necessarily exactly the same (Barausse et al. 2017). Although changing ingredients such the employed scaling relation should not change the occurrence of mergers due to triple interactions (our triplet-induced merger fractions are fairly independent on the mass scale of the problem, see Paper II), things can be different for MBH evolution models relying on radically different merger histories. For example, if galaxy merger rates (which are relatively poorly constrained by observations) are much less frequent, then the occurrence of subsequent mergers is much rarer, implying a lower triplet formation rate. This, in turn, will cause a larger suppression of the GWB. We note, however, that in *Model-stalled* the majority of mergers actually involve quadruple systems. This means that mergers are frequent enough that, in *Model-stalled*, the vast majority of galaxies hosts a MBHB *at any time* along cosmic history. In practice, only a radically different structure formation

scenario in which massive galaxies experience on average less than one merger along the cosmic history would result in a larger suppression of the GW signal in *Model-stalled* compared to *Model-delayed*. In such scenario, in fact, the paucity of galaxy interactions would imply a very low probability for a given galaxy to experience two subsequent mergers, and MBH triplets would not form. We note, however, that both semianalytic and numerical simulations support a rich merger history for massive galaxies (see, e.g. Volonteri, Haardt & Madau 2003; De Lucia & Blaizot 2007; Oser et al. 2010; Kelley, Blecha & Hernquist 2017), making the aforementioned scenario extremely unlikely. In support to this consideration, we note that by employing a completely different framework and MBH evolution model, Ryu et al. (2017) find a very similar suppression in the GWB normalisation of about 30%.

Although these are substantial caveats, we argue that they are unlikely to strongly influence the results obtained here and our conclusions are thus robust.

6 CONCLUSIONS

We have explored the effect of MBH triple interactions on the GW signal produced by a cosmic population of MBHs. To this purpose, we have coupled a large library of numerical simulations of triple interactions (Paper I, Paper II) to a semianalytic model for galaxy- and MBH-evolution (Barausse 2012). The numerical simulations solve the 3-body equations of motion consistently derived from the 3-body PN Hamiltonian (Paper I) through 2.5PN order, and include the effect of the galactic potential and an analytic treatment of dynamical friction and stellar hardening. The library of outcomes is then implemented within the semianalytic model, which keeps track of the evolution of individual MBHs and the formation of MBH binaries, triplets and quadruplets following galaxy mergers. This framework has allowed us to assess the effect of triple (and quadruple) interactions on the MBHB cosmic merger rate and on the expected GWB in the PTA band.

In particular, we have considered two models for the dynamics of MBHBs. In our fiducial model (labelled *Model-delayed*) MBHBs merge on timescales of millions-to-billions years, consistently estimated from the properties of the host galaxy. In the PTA band, most mergers occur in gas-poor galaxies and typical merger timescales due to stellar-driven hardening (assuming efficient loss cone replenishment) are of the order of Gyrs. This allows the formation of several triple MBH systems due to subsequent galaxy mergers, but the formation of 'standard' MBHBs is still the dominant coalescence channel. We have then considered an extreme model where all standard dynamical processes are inefficient at driving MBHBs to sub-pc scales, and MBHBs stall close to their hardening radius (*Model-stalled*). The rationale behind this model was to investigate the outcome of the most pessimistic scenario from the GW generation standpoint; naively, if all MBHBs stall, no (or very little, as pointed out by DB17) GW signal is expected in the PTA band. However, mergers can still be triggered by triple (and quadruple) MBH interactions following subsequent galaxy mergers, a possibility that was not accounted for in DB17 and which we have explored here in depth.

Our main results can be summarised as follows:

(i) Even if the final-parsec problem is not naturally solved by the interaction of MBHBs with their stellar and gaseous environment (i.e. in *Model-stalled*), triple interactions can still lead a large number of MBHBs to final coalescence. In the specific galaxy and MBH evolution model explored here, when stalling is assumed, the MBHB merger rate is only suppressed by a factor of ≈ 4 (cf table 1) in the mass relevant to PTA observations. Those mergers are the result of triple interactions.

(ii) The implied GWB background is only suppressed by a factor of about 2-to-3 in the relevant PTA frequency range $1\ \mathrm{nHz} < f < 10\ \mathrm{nHz}$.

(iii) Triple MBH interactions naturally produce eccentric binaries. This causes the GWB to be generally flatter than the standard $f^{-2/3}$ power law. However, no clear turnover is seen, at least above $\gtrsim 0.1$ nHz, due to the wide range of eccentricities of the binaries.

(iv) The most massive MBHBs $\mathcal{M} > 10^{8.5}\mathrm{M}_{\odot}$, which are the most likely to be individually resolved, can have eccentricities larger than 0.9 at the relevant PTA frequencies. Still, the majority of them tends to be circular or mildly eccentric.

(v) When coupling the GWB suppression due to stalling to a pessimistic MBHB population model that predicts a particularly low GWB, we still obtain amplitude normalisations at the level $A \gtrsim 10^{-16}$.

(vi) The predicted amplitude is well within the reach of SKA. We find that a putative array monitoring 50 pulsars at 100 ns level has a 90% chance of detection after 15 years of observation. In general, we find that stalling will delay GWB detection by only about 3-to-6 years depending on the number of pulsars in the array.

A particularly relevant result is that signal amplitudes below $A \approx 10^{-16}$ are extremely unlikely even in the most pessimistic scenario in which (i) MBHs are intrinsically less massive than predicted by standard MBH-host galaxy relations *and* (ii) MBHBs stall.

Our main claim is therefore that, because of triple interactions, stalling does not strongly decrease the level of the GWB in the PTA frequency range. The only other way to pose a threat to future PTA detections is if the opposite of stalling is realised in Nature; i.e. if an extremely efficient coupling with the environment swiftly drives MBHBs through the PTA band, which would cause a low-frequency turnover in the GWB. To be dangerous, such turnover should be at frequencies well above 10 nHz, which for realistic environments is never the case, unless *all* MBHBs are extremely ($e > 0.99$) eccentric. This seems a very unlikely possibility since simulations of MBHBs in stellar environments generally find a *range* of eccentricities $0 \lesssim e \lesssim 1$. The most important implication is that with the advent of MeerKAT, FAST and SKA, PTAs *will* detect a GW signal from merging MBHBs, provided that those instruments bring an-order-of-magnitude improvement over current PTA sensitivities.

Our results are subject to a number of caveats that we have extensively discussed: a very approximate treatment of quadruple interactions, our assumption that MBHBs are driven to separations comparable with their influence radius on timescales shorter than the Hubble time, the direct application of our findings to different MBHB populations to

derive a lower limit for the expected GWB amplitude. We have argued that none of those caveats is critical and that the results presented here are robust. Even if everything conspires to produce the lowest possible GWB amplitude, a typical SKA-based PTA will still have a $> 90\%$ probability of detecting a signal within 15 years of data collection, which strengthens the scientific case of this observatory and which is a good reason to look with optimism at the future of GW astrophysics in the nHz band.

ACKNOWLEDGEMENTS

MB and FH acknowledge partial financial support from the INFN TEONGRAV specific initiative. MB acknowledges the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support. This work was supported by the H2020-MSCA-RISE-2015 Grant No. StronGrHEP-690904. AS is supported by a University Research Fellowship of the Royal Society. This work has made use of the Horizon Cluster, hosted by the Institut d'Astrophysique de Paris. We thank Stephane Rouberol for running smoothly this cluster for us.

REFERENCES

Antonini F., Barausse E., Silk J., 2015a, ApJ, 812, 72
Antonini F., Barausse E., Silk J., 2015b, ApJ, 806, L8
Arzoumanian Z. et al., 2016, ApJ, 821, 13
Barausse E., 2012, MNRAS, 423, 2533
Barausse E., Shankar F., Bernardi M., Dubois Y., Sheth R. K., 2017, MNRAS, 468, 4782
Begelman M. C., Blandford R. D., Rees M. J., 1980, Nature, 287, 307
Blecha L. et al., 2016, MNRAS, 456, 961
Bonetti M., Haardt F., Sesana A., Barausse E., 2016, MNRAS, 461, 4419
Bonetti M., Haardt F., Sesana A., Barausse E., 2017, ArXiv e-prints
Boylan-Kolchin M., Ma C.-P., Quataert E., 2008, MNRAS, 383, 93
Bromley B. C., Kenyon S. J., Geller M. J., Barcikowski E., Brown W. R., Kurtz M. J., 2006, ApJ, 653, 1194
Cattaneo A., Dekel A., Devriendt J., Guiderdoni B., Blaizot J., 2006, MNRAS, 370, 1651
Chen S., Sesana A., Del Pozzo W., 2017, MNRAS, 470, 1738
Cuadra J., Armitage P. J., Alexander R. D., Begelman M. C., 2009, MNRAS, 393, 1423
De Lucia G., Blaizot J., 2007, MNRAS, 375, 2
Dekel A., Birnboim Y., 2006, MNRAS, 368, 2
Dekel A. et al., 2009, Nature, 457, 451
Dekel A., Sari R., Ceverino D., 2009, ApJ, 703, 785
Desvignes G. et al., 2016, MNRAS, 458, 3341
Dosopoulou F., Antonini F., 2017, ApJ, 840, 31
Dotti M., Sesana A., Decarli R., 2012, Advances in Astronomy, 2012, 940568
Dvorkin I., Barausse E., 2017, ArXiv e-prints
Ferrarese L., Merritt D., 2000, ApJ, 539, L9
Foster R. S., Backer D. C., 1990, ApJ, 361, 300
Gebhardt K. et al., 2000, ApJ, 539, L13
Granato G. L., De Zotti G., Silva L., Bressan A., Danese L., 2004, ApJ, 600, 580
Heger A., Woosley S. E., 2002, ApJ, 567, 532
Hellings R. W., Downs G. S., 1983, ApJ, 265, L39
Hobbs G. et al., 2010, Classical and Quantum Gravity, 27, 084013
Hoffman L., Loeb A., 2007, MNRAS, 377, 957
Hogg D. W., 1999, ArXiv Astrophysics e-prints
Iwasawa M., Funato Y., Makino J., 2006, ApJ, 651, 1059
Jaffe A. H., Backer D. C., 2003, ApJ, 583, 616
Janssen G. et al., 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 37
Jenet F. A. et al., 2006, ApJ, 653, 1571
Kauffmann G., Haehnelt M., 2000, MNRAS, 311, 576
Kelley L. Z., Blecha L., Hernquist L., 2017, MNRAS, 464, 3131
Kelley L. Z., Blecha L., Hernquist L., Sesana A., Taylor S. R., 2017, MNRAS, 471, 4508
Khan F. M., Preto M., Berczik P., Berentzen I., Just A., Spurzem R., 2012, ApJ, 749, 147
Klein A. et al., 2016, Phys. Rev. D, 93, 024003
Kormendy J., Richstone D., 1995, ARA&A, 33, 581
Kozai Y., 1962, AJ, 67, 591
Kulier A., Ostriker J. P., Natarajan P., Lackner C. N., Cen R., 2015, ApJ, 799, 178
Kulkarni G., Loeb A., 2012, MNRAS, 422, 1306
Lapi A., Raimundo S., Aversa R., Cai Z.-Y., Negrello M., Celotti A., De Zotti G., Danese L., 2014, ApJ, 782, 69
Lentati L. et al., 2015, MNRAS, 453, 2576
Lidov M. L., 1962, Planet. Space Sci., 9, 719
MacFadyen A. I., Milosavljević M., 2008, ApJ, 672, 83
Madau P., Haardt F., Dotti M., 2014, ApJ, 784, L38
Madau P., Rees M. J., 2001, ApJ, 551, L27
Magorrian J. et al., 1998, AJ, 115, 2285
McWilliams S. T., Ostriker J. P., Pretorius F., 2014, ApJ, 789, 156
Nixon C. J., Cossins P. J., King A. R., Pringle J. E., 2011, MNRAS, 412, 1591
Oser L., Ostriker J. P., Naab T., Johansson P. H., Burkert A., 2010, ApJ, 725, 2312
Parkinson H., Cole S., Helly J., 2008, MNRAS, 383, 557
Peters P. C., Mathews J., 1963, Phys. Rev., 131, 435
Phinney E. S., 2001, ArXiv Astrophysics e-prints
Press W. H., Schechter P., 1974, ApJ, 187, 425
Quinlan G. D., 1996, New A, 1, 35
Rajagopal M., Romani R. W., 1995, ApJ, 446, 543
Ravi V., Wyithe J. S. B., Hobbs G., Shannon R. M., Manchester R. N., Yardley D. R. B., Keith M. J., 2012, ApJ, 761, 84
Ravi V., Wyithe J. S. B., Shannon R. M., Hobbs G., 2015, MNRAS, 447, 2772
Reardon D. J. et al., 2016, MNRAS, 455, 1751
Rodriguez-Gomez V. et al., 2015, MNRAS, 449, 49
Rosado P. A., Sesana A., Gair J., 2015, MNRAS, 451, 2417
Ryu T., Perna R., Haiman Z., Ostriker J. P., Stone N. C., 2017, ArXiv e-prints
Sazhin M. V., 1978, Soviet Ast., 22, 36
Sesana A., 2013, MNRAS, 433, L1
Sesana A., 2015, in Astrophysics and Space Science Proceedings, Vol. 40, Gravitational Wave Astrophysics, Sopuerta C. F., ed., p. 147

Sesana A., Barausse E., Dotti M., Rossi E. M., 2014, ApJ, 794, 104
Sesana A., Haardt F., Madau P., 2006, ApJ, 651, 392
Sesana A., Haardt F., Madau P., Volonteri M., 2004, ApJ, 611, 623
Sesana A., Khan F. M., 2015, MNRAS, 454, L66
Sesana A., Shankar F., Bernardi M., Sheth R. K., 2016, MNRAS, 463, L6
Sesana A., Vecchio A., Colacino C. N., 2008, MNRAS, 390, 192
Sesana A., Vecchio A., Volonteri M., 2009, MNRAS, 394, 2255
Shankar F. et al., 2016, MNRAS, 460, 3119
Shannon R. M. et al., 2015, Science, 349, 1522
Siemens X., Ellis J., Jenet F., Romano J. D., 2013, Classical and Quantum Gravity, 30, 224015
Taffoni G., Mayer L., Colpi M., Governato F., 2003, MNRAS, 341, 434
Taylor J. H., 1992, Philosophical Transactions of the Royal Society of London Series A, 341, 117
Taylor S. R., Huerta E. A., Gair J. R., McWilliams S. T., 2016a, ApJ, 817, 70
Taylor S. R., Simon J., Sampson L., 2017, Physical Review Letters, 118, 181102
Taylor S. R., Vallisneri M., Ellis J. A., Mingarelli C. M. F., Lazio T. J. W., van Haasteren R., 2016b, ApJ, 819, L6
The NANOGrav Collaboration et al., 2015, ApJ, 813, 65
Vasiliev E., Antonini F., Merritt D., 2015, ApJ, 810, 49
Verbiest J. P. W. et al., 2016, MNRAS, 458, 1267
Volonteri M., Haardt F., Madau P., 2003, ApJ, 582, 559
Volonteri M., Lodato G., Natarajan P., 2008, MNRAS, 383, 1079
Wyithe J. S. B., Loeb A., 2003, ApJ, 590, 691