Publications of the Astronomical Society of Australia (PASA), Vol. 33, e012, 11 pages (2016).

doi:10.1017/pasa.2016.12

The Evolution of Self-Gravitating Accretion Discs

Ken Rice[1,2]

[1]SUPA, Institute for Astronomy, Royal Observatory, Blackford Hill, Edinburgh, EH93HJ, UK
[2]Email: wkmr@roe.ac.uk

(Received November 17, 2015; Accepted February 26, 2016)

Abstract

It is quite likely that self-gravity will play an important role in the evolution of accretion discs, in particular, those around young stars, and those around supermassive black holes. We summarise, here, our current understanding of the evolution of such discs, focussing more on discs in young stellar system, than on discs in active galactic nuclei. We consider the conditions under which such discs may fragment to form bound objects, and when they might, instead, be expected to settle into a quasi-steady, self-regulated state. We also discuss how this understanding may depend on the mass of the disc relative to the mass of the central object, and how it might depend on the presence of external irradiation. Additionally, we consider whether or not fragmentation might be stochastic, where we might expect it to occur in an actual protostellar disc, and if there is any evidence for fragmentation actually playing a role in the formation of planetary-mass bodies. Although there are still a number of outstanding issue, such as the convergence of simulations of self-gravitating discs, whether or not there is more than one mode of fragmentation, and quite what role self-gravitating discs may play in the planet-formation process, our general understanding of these systems seems quite robust.

Keywords: accretion, accretion discs – instabilities – planets and satellites: formation – protoplanetary disks – stars: formation

1 INTRODUCTION

It is likely that self-gravity will, at times, play an important role in the evolution of accretion discs. In particular, we expect it to be important in discs around active galactic nuclei (Nayakshin, Cuadra, & Springel 2007) and in discs around very young protostars (Lin & Pringle 1987). Although we will discuss some general properties of self-gravitating discs, we will focus more on protostellar discs, than on other types of potentially self-gravitating accretion discs.

An accretion disc will be self-gravitating, or susceptible to the growth of the gravitational instability, if the Toomre parameter, Q, is close to unity (Safronov 1960; Toomre 1964)

$$Q = \frac{c_s \Omega}{\pi G \Sigma} \sim 1, \qquad (1)$$

where c_s is the sound speed, Ω is the epicyclic frequency (equal to the orbital angular frequency in Keplerian discs), Σ is the disc's surface density, and G is the gravitational constant. This parameter, however, only tells us if the disc's self-gravity is important and if it will likely develop a gravitational instability. It doesn't, however, tell is how such discs will actually evolve.

There are two likely pathways for a self-gravitating disc. It could settle into a quasi-steady, self-regulated state (Paczynski 1978), or it could become so unstable that it fragments into bound objects (Boss 1998). The latter could be a potential star-formation process in discs around active galactic nuclei (Levin & Beloborodov 2003), or a mechanism for forming planets, or brown dwarfs, in discs around young stars (Boss 2000). Even if a disc does not fragment, a self-gravitating phase could play a very important role in driving angular momentum transport and, hence, allowing mass to accrete onto the central protostar (Lin & Pringle 1987; Rice, Mayo, & Armitage 2010).

In this paper, we will summarise our current understanding of the evolution of self-gravitating accretion discs. There are a number of factors that likely determine the evolution of such discs. In particular, the balance between heating and cooling, and the mass of the disc relative to the mass of the central object. We'll discuss if we can establish the conditions which determine whether a self-gravitating disc will settle into a quasi-steady state, or fragment, and—if so—when such conditions are valid, and when they might be violated. We'll look at the role of external factors, such as external irradiation and perturbations from stellar, or sub-stellar, companions. We'll also discuss, in the context of protostellar discs, whether or not we expect fragmentation to actually occur and—if so—where it is likely to occur and what we might expect the outcome to be.

The paper is structured in the following way. In Section 2, we focus on the quasi-steady evolution of self-gravitating discs, how such discs evolve under the influence of external irradiation, what happens if the disc mass is high relative to the mass of the central object, and what might happen if we consider more realistic scenarios. In Section 3, we focus on the fragmentation of self-gravitating discs. In particular, we consider what conditions determine if a disc will fragment, when such conditions may apply and when they may be violated, could it be stochastic rather than precisely determined, and could fragmentation be triggered by external perturbations. In Section 4, we then discuss these results and draw conclusions.

2 QUASI-STEADY EVOLUTION

As already mentioned, a self-gravitating disc is one that has sufficient mass, or that is cold enough, so that the Toomre parameter, Q, is of order unity. Such a self-gravitating disc will generate an instability that will dissipate and heat the disc. If the cooling rate can match the rate at which the instability is heating the disc, the system may then sustain a state of marginal stability (Paczynski 1978). If not, the system may fragment to form bound objects, potentially protoplanets or brown dwarfs in discs around young stars (Boss 2000), or stars in discs around active galactic nuclei (Levin & Beloborodov 2003; Bonnell & Rice 2008).

If such a system does maintain a state of marginal stability, one would expect the instability to transport angular momentum. A standard way to describe angular momentum transport in discs is to assume that it is viscous and can be described using the α-prescription (Shakura & Sunyaev 1973). Energy transport in a self-gravitating disc, however, is not diffusive and so doesn't, strictly speaking, satisfy this condition (Balbus & Papaloizou 1999). Numerical studies (Laughlin & Rozyczka 1996; Nelson, Benz, & Ruzmaikina 2000; Pickett et al. 2000) have, however, shown that the α-prescription can be a reasonable way to parameterise angular momentum transport in self-gravitating discs.

One of the first studies to quantify this (Gammie 2001) used a local shearing sheet in which the cooling time, τ, was assumed to be a constant, β, divided by the local angular frequency, Ω:

$$\tau = \beta\Omega^{-1}. \tag{2}$$

Such a cooling formalism has now become known as β-cooling. What this work showed was that, as expected, if the disc settled into a quasi-steady state, the dissipation rate would match the imposed cooling, and the effective viscous α would then be given by

$$\alpha = \frac{4}{9\gamma(\gamma - 1)\beta}, \tag{3}$$

where γ is the specific heat ratio. Figure 1 shows an example of the quasi-steady, surface density structure in a local shearing sheet simulation in which the imposed cooling is assumed to have $\beta = 10$ and in which $\gamma = 2$. Figure 2 shows the resulting α evolution. The simulation is initially run with no cooling, which is then turned on at $t = 50$. There is an initial burst and then the system settles into a quasi-steady state with an approximately constant α. The dashed line shows the α value to which we'd expect it to settle, based on Equation (3).

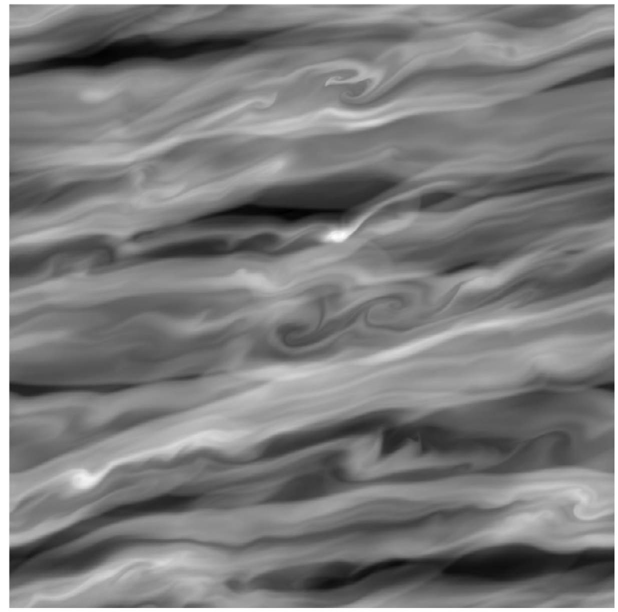

Figure 1. A snapshot showing the surface density structure in a shearing sheet simulation with $\beta = 10$ and $\alpha = 2$. The disc has settled into a quasi-steady. self-regulated state in which heating via dissipation of the gravitational instability is balancing the imposed cooling.

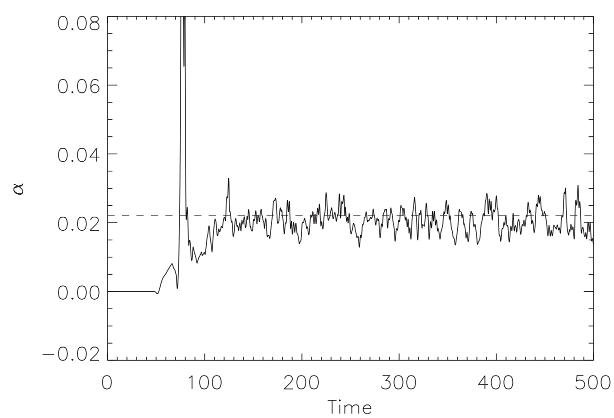

Figure 2. The effective α plotted against time for the simulation shown in Figure 1. Cooling is turned on at $t = 50$, after which there is a sudden burst. However, by $t = 100$, the simulation has settled into a quasi-steady state with a roughly constant α, consistent with what is expected from energy balance (dashed line).

Other studies employing different types of numerical methods, such as Smoothed particle hydrodynamics (SPH, Lodato & Rice 2004; Rice, Lodato, & Armitage 2005) and grid-based methods (Mejía et al. 2005), have similarly shown that a self-gravitating disc will tend to settle to a quasi-steady state in which Q is close to unity and in which angular momentum transport can be described using an α-formalism in which α is given by Equation (3).

There are, however, some caveats to the above that we will discuss in more detail later. It applies in situations where the disc mass is low enough that the local approximation is valid, and also requires that the cooling time is not so short that the instability becomes non-linear and the disc breaks up into fragments.

2.1. Irradiation

A fundamental assumption in some of the early work on the evolution of self-gravitating discs (e.g., Gammie 2001) was that there was only a single heating source; dissipation of the gravitational instabilty. In such a scenario, the cooling balances only this heating source and the effective α is given simply by Equation (3). In the presence of an additional heating source (such as external irradiation), it is more complex, since—in a state of marginal stability— the cooling balances both heating due to the instability and the additional heating source.

For a given cooling time, the effective α is reduced in the presence of an additional heating source, compared to a situation where the only heating source is the gravitational instability itself. It can be shown Rice et al. (2011) that, in the presence of external irradiation, one can approximate the effective α using

$$\alpha = \frac{4}{9\gamma(\gamma-1)\beta}\left(1 - \frac{\langle\Sigma\rangle c_{\rm so}^2}{\langle\Sigma c_{\rm s}^2\rangle}\right), \tag{4}$$

where Σ is the local surface density, $c_{\rm so}$ is the sound speed to which the disc would settle in the presence of external irradiation only, and $c_{\rm s}$ is the sound speed to which the disc settles if it does tend to a state of marginal stability. This is similar to the form suggested by Kratter, Murray-Clay, & Youdin (2010). This can then be recast as

$$\alpha = \frac{4}{9\gamma(\gamma-1)\beta}\left(1 - \frac{Q_{\rm irr}^2}{Q_{\rm sat}^2}\right), \tag{5}$$

where $Q_{\rm irr}$ is the Q value to which it would settle in the presence of external irradiation only, while $Q_{\rm sat}$ is the saturated Q value to which the disc settles in the presence of both the external irradiation and heating through dissipating the gravitational instability. Using shearing sheet simulations, Rice et al. (2011) showed that this relationship is indeed reasonable, and the results are illustrated in Figure 3 (from Rice et al. 2011) which shows how the effective α varies with increasing levels of external irradiation, when the underlying cooling timescale is $\beta = 9$.

The important point here is that if there are other sources of heating, the strength of the instability is not set by the cooling timescale, β, alone. In the presence of other heating sources, the strength of the instability, and the magnitude of α, will be smaller than if the only heating source is the gravitational instability.

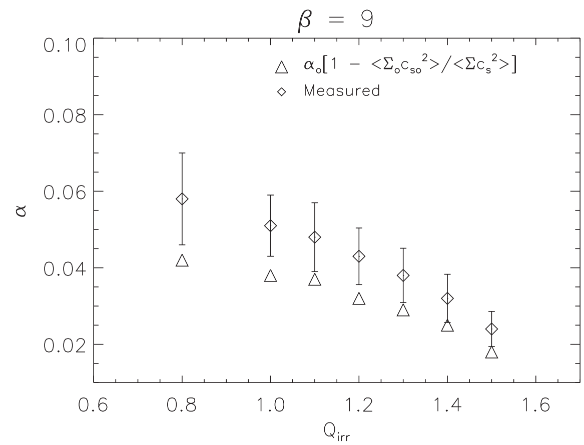

Figure 3. Figure showing how the quasi-steady α value varies with the level external irradiation (taken from Rice et al. 2011). The triangles are an initial semi-analytic estimate, while the diamonds with error bars are from the shearing sheet simulations. Although there is some discrepancy, it's clear that the effective α decreases with increasing levels of external irradiation.

2.2. Local vs non-local

Another assumption underlying the idea that angular momentum transport via the gravitational instability can be treated as essentially viscous, is that the shear stress (or, equivalently, the viscous α) depends primarily on the local conditions in the disc. Gravity is, however, inherently global, and so such an assumption isn't strictly correct. Balbus & Papaloizou (1999) have shown that the energy equation for a self-gravitating disc has extra terms associated with global wave transport. They do, however, suggest that self-gravitating discs that hover near marginal gravitational stability may behave in a manner that is well described by a local α model.

Early global simulations (Laughlin & Bodenheimer 1994) seemed to indicate that simple α models could reproduce the density evolution of self-gravitating accretion discs. Such simulations, however, initially ignored heating and cooling processes in the disc, which play a fundamental role in setting the quasi-steady nature of such systems (Gammie 2001). Imposing a β-cooling formalism in global disc simulations, Lodato & Rice (2004) illustrated that quasi-steady, self-gravitating discs do indeed behave like α discs and that the local model is a reasonable approximation.

The local nature of self-gravitating angular momentum transport does, however, depend on the mass of the disc relative to the mass of the central object. As the disc-to-star mass ratio increases, the disc becomes thicker and global effects are likely to become more important. Lodato & Rice (2004) only considered discs that had masses less than one quarter the mass of the central star. Later work (Lodato & Rice 2005) considered much higher mass ratios. What was found was that as the disc-to-star mass ratio increases, lower order spiral modes start to dominate. For mass ratios of ~0.5, a transient $m = 2$ spiral develops that rapidly redistributes mass and then allows the disc settle into a quasi-steady, self-regulated state. For mass ratios approaching unity, however,

doi:10.1017/pasa.2016.12

the disc never seems to settle into some kind of quasi-steady, self-regulated state. However, it does appear to satisfy the basic criteria in a time-averaged sense.

The basic picture therefore appears to be that for mass ratios below ∼0.5 a self-gravitating disc will rapidly settle into a quasi-steady, self-regulated state in which $Q \sim 1$, heating and cooling are in balance, and in which angular momentum transport is pseudo-viscous and can be described by an effective α. As the mass ratio increases, lower m-modes begin to dominate, producing global spiral density waves that either rapidly redistribute mass, allowing the disc to then settle into a quasi-steady, self-regulated state, or—for sufficiently high mass ratios—that have time varying amplitudes leading to the disc only being quasi-steady in a time averaged sense.

2.3. Realistic cooling

A great deal of work on the evolution of self-gravitating accretion discs has used the basic β-formalism when implementing cooling. This, however, is clearly not a realistic representation of how such discs will actually cool, and was never intended to be so. It is simply a basic way in which to represent cooling so as to investigate how such discs evolve under a range of different cooling scenarios.

It's clear that in reality, a self-gravitating disc will not have a cooling function that can be described via a single β parameter (Pickett et al. 2003) and that the actual evolution may depend on the form of the cooling function (Mejía et al. 2005). A number of studies have indeed used more realistic cooling formalisms. For example, Boss (2001, 2003) and Mayer et al. (2007) used simulations with radiative transfer to understand the fragmentation of self-gravitating discs, Cai et al. (2006) considered how the disc metallicity might influence its evolution, Boley et al. (2006) considered the quasi-steady evolution of discs with realistic cooling and realistic opacities, and Meru & Bate (2010) probed the parameter space where disc fragmentation may occur, using three-dimensional hydrodynamical simulations with radiative transfer.

More recently, Forgan et al. (2011) also considered the evolution of self-gravitating discs using a more realistic cooling formalism, but with the goal of trying to understand its quasi-steady evolution and, in particular, if the local approximation still applies when the cooling is more realistic. They performed three-dimensional SPH simulations of discs around young stars in which a radiative transfer formalism (Stamatellos et al. 2007; Forgan et al. 2009) was used to estimate how the disc cooled.

Their results suggest that for discs with masses below about half that of the central star, the general behaviour is well described by the local approximation. An estimate of the local cooling time, when $Q \sim 1$, can be used to determine the local value of α. However, as the mass ratio increases, the behaviour becomes more global, and the local cooling time is not a good indicator of the effective α, except in a time-averaged sense.

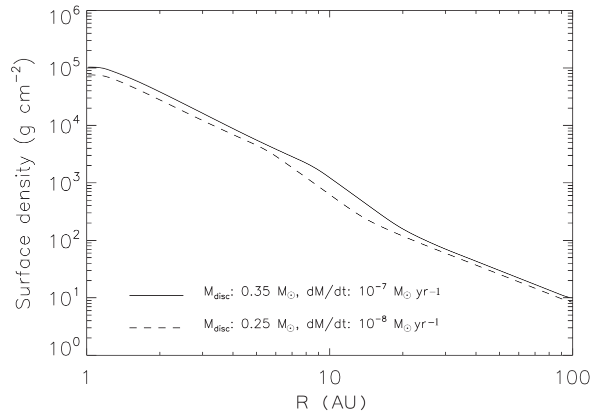

Figure 4. Examples of two surface density profiles for self-gravitating discs that are in a quasi-steady state. In these examples, there is no external irradiation and standard opacities are assumed. For given opacities, the profile is essentially unique for a specific mass accretion rate. The mass accretion rates here are $\sim 10^{-7}$ M$_\odot$ yr^{-1} (solid line) and $\sim 10^{-8}$ M$_\odot$ yr^{-1} (dashed line) and the disc masses are ∼0.35 M$_\odot$ (solid line) and ∼0.25 M$_\odot$ (dashed line). This illustrates that a relatively small change in disc mass can produce a substantial change in accretion rate, indicating that disc self-gravity is only likely to play an important role in angular momentum transport when these systems are very young, and the disc mass is relatively high.

Given the above, it is possible to use semi-analytic calculations to determine the realistic structure of quasi-steady, self-gravitating accretion discs (Rafikov 2009; Clarke 2009; Rice & Armitage 2009), at least for those with masses less than about half that of the central star. For example, for a given opacity one can determine the cooling rate for a given surface density and, hence, the local effective viscosity (α). Using standard one-dimensional accretion disc models (Pringle 1981), one can then determine the surface density profile for a disc accreting at a specified accretion rate.

For example, Figure 4 shows the surface density profile of a quasi-steady, self-gravitating disc around a 1 M$_\odot$ star accreting at $\sim 10^{-7}$ M$_\odot$ yr^{-1} (solid line) and accreting at $\sim 10^{-8}$ M$_\odot$ yr^{-1} (dashed line), for standard opacities (Bell & Lin 1994) and in the absence of external irradiation. Given the opacities, the profile is essentially unique for a specific mass accretion rate. What Figure 4 also shows is that there is quite a strong mass dependence. Assuming an outer disc radius of 100 AU, the disc mass of the disc accreting at $\sim 10^{-7}$ M$_\odot$ yr^{-1} is ∼0.35 M$_\odot$, while that of the disc accreting at $\sim 10^{-8}$ M$_\odot$ yr^{-1} is ∼0.25 M$_\odot$. Less than a factor of two change in disc mass, produces about an order of magnitude change in accretion rate.

This has a number of consequences. To explain TTauri-like, and higher, accretion rates, self-gravitating discs need to be quite massive, relative to the mass of the central star. Given that disc lifetimes are typically 5 Myr, or less (Haisch, Lada, & Lada 2001), self-gravity is unlikely to explain accretion rates during the later stages of a discs lifetime, since it would imply much longer lifetimes than those observed.

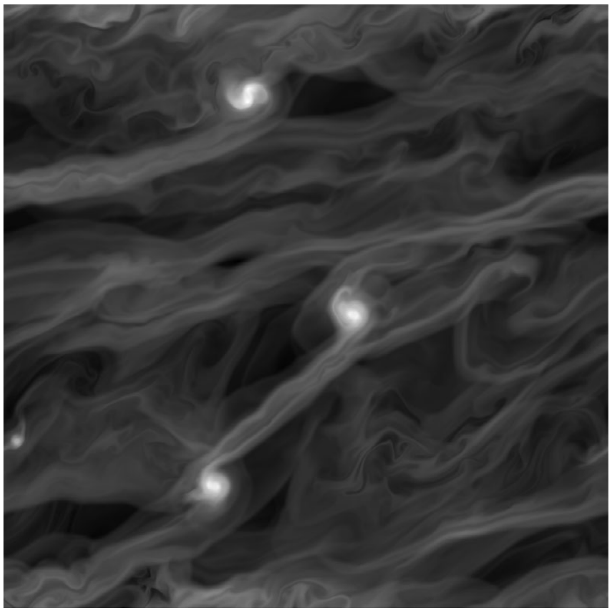

Figure 5. A snapshot showing the surface density structure in a shearing sheet simulation with $\beta = 2$ and $\gamma = 2$. There are a number of dense clumps, indicating that the rapid cooling in this simulation has led to the disc fragmenting, rather than settling into a quasi-steady, self-regulated state.

Similarly, observed spiral structures in protostellar discs are unlikely to be due to disc self-gravity unless the disc mass is still relatively high (Dong et al. 2015). For discs with masses less than $\sim$0.1 $M_\odot$, the instability is likely to be weak, is unlikely to generate the stresses that can dominate the angular momentum transport process, and is unlikely to have spiral density waves that are sufficiently strong so as to be observable.

3 FRAGMENTATION

The previous section focussed on the quasi-steady evolution of self-gravitating accretion discs. For sufficiently unstable discs, however, one possible outcome is that they fragment to form bound objects. It has been suggested that these could be protoplanets in discs around young stars (Boss 1998), or protostars in discs around supermassive black holes (Levin & Beloborodov 2003).

Again using shearing sheet simulations, Gammie (2001) showed that a disc will fragment if the cooling time is sufficiently fast; in this case, the fragmentation boundary being at $\beta = 3$. This is illustrated in Figure 5 which shows the surface density structure in a shearing sheet simulation with $\beta = 2$ and $\gamma = 2$. It is clear that there are a number of dense clumps forming in the disc.

A similar fragmentation boundary was obtained using three-dimensional, global simulations (Rice et al. 2003). However, as Equation (3) shows, the effective α depends on the specific heat capacity γ. Consequently, it was shown that the fragmentation boundary is actually set by a maximum α that can be sustained in a quasi-steady, self-regulated disc, not by the cooling time specifically (Rice et al. 2005). For example, the cooling time for fragmentation if $\gamma = 2$ is smaller than for $\gamma = 1.4$.

Additionally, one has to be careful because all of this early work was done under the assumption that dissipation of the gravitational instability was the only heating source. As discussed above, the presence of external irradiation weakens the instability and reduces the effective α (Rice et al. 2011). Therefore, it is more reasonable to regard the fragmentation boundary as being associated with the maximum self-gravitating α that a disc can sustain in a self-regulated, quasi-steady state, than as the minimum cooling time that can be imposed without the disc fragmenting.

Furthermore, the above analyses have all used simulations in which the local approximation is reasonably valid. As discussed above, as the disc becomes more massive (relative to the mass of the central object) global modes start to dominate, and the local approximations are only valid in a time averaged sense. When global modes start to dominate, it appears that a self-gravitating disc can sustain larger α values without fragmenting, than would be expected based on the local approximation only (Lodato & Rice 2005). Simulations of collapsing cloud cores with radiative transfer also indicate that the behaviour is quite different to what would be expected based on the local approximation (Forgan et al. 2011; Tsukamoto et al. 2015). Discs that don't fragment can have α values considerably higher than the local approximation would suggest is possible. As Forgan et al. (2011) illustrates, however, the Reynolds stress, driven by velocity shear from material from the infalling envelope contacting the disc, dominates, and the gravitational component of the stress still typically produces α values less than 0.1.

3.1. Perturbation amplitudes

If we focus again on the case where the local approximation is valid, one way to understand fragmentation is to consider the relationship between α and the perturbation amplitudes. Cossins, Lodato, & Clarke (2009) show that the transport of energy and angular momentum, by the spiral waves driven by the gravitational instability, depends on the surface density perturbation $\delta\Sigma$. Since the angular momentum transport can be described via an effective α, which itself depends on the cooling timescale β, Cossins et al. (2009) show that the perturbation amplitudes depend on β through

$$\frac{\langle \delta\Sigma \rangle}{\langle \Sigma \rangle} \simeq \frac{1.0}{\sqrt{\beta}}. \quad (6)$$

Similarly, you can cast this as (Rice et al. 2011)

$$\frac{\langle \delta\Sigma \rangle}{\langle \Sigma \rangle} \simeq \sqrt{\alpha}. \quad (7)$$

Figure 6 (taken from Rice et al. 2011) shows a plot of time-averaged α against $\langle \delta\Sigma \rangle / \langle \Sigma \rangle$ illustrating that as α increases the perturbation amplitudes increase. Figure 6 also shows

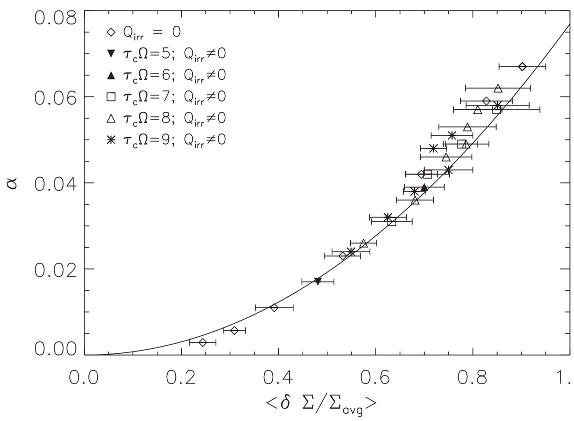

Figure 6. A figure showing time-averaged α plotted against the mean perturbation amplitude ($\langle \delta \Sigma \rangle / \langle \Sigma \rangle$), for a variety of different shearing sheet simulations. It's clear that there is a relationship between α and the perturbation amplitudes, with the perturbation amplitudes increasing with increasing α. In the absence of external irradiation ($Q_{\rm irr} = 0$), this would also indicate a relationship between the perturbation amplitudes and β (Cossins et al. 2009), but in the presence of external irradiation ($Q_{\rm irr} \neq 0$) it is α that sets the perturbation amplitude, not simply β. (Figure from Rice et al. 2011.)

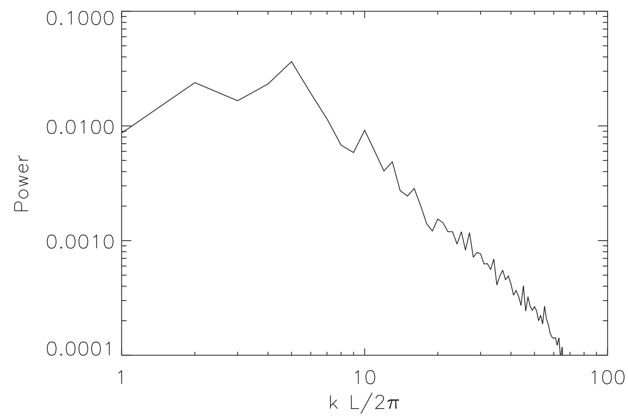

Figure 7. Figure showing the power spectrum of the perturbations in a shearing sheet simulation with $\beta = 10$. The length of each side of the sheet is L, and so this figure shows that most of the power is at scales a few times smaller than the sheet size and that there is very little power at very small scales.

results from simulations with no external irradiation ($Q_{\rm irr} = 0$) and with external irradiation ($Q_{\rm irr} \neq 0$) illustrating that it is α that determines the perturbation amplitudes, not simply the cooling time β.

Essentially, the basic picture is that as the self-gravitating α increases, the perturbations become larger and, there will be an α value beyond which the perturbations collapse and the disc fragments. Most simulations suggest that this occurs at an α value of about 0.06 (Rice et al. 2005), but this should probably not be seen as some kind of hard boundary, as we'll discuss in more detail later.

3.2. Convergence

Recently, it has been suggested that the fragmentation boundary does not converge to a fixed α value as the numerical resolution of the simulation increases (Meru & Bate 2011, 2012; Baehr & Klahr 2015). There have been a number of attempts to explain this. Lodato & Clarke (2011) and Meru & Bate (2012) suggest that numerical viscosity may play a larger role than previously thought (e.g., Murray 1996) and that, consequently, simulations will need to consider higher resolutions than have been used in the past. Paardekooper, Baruteau, & Meru (2011) show that the lack of convergence could be related to edge effects in simulations with very smooth initial conditions. This apparent non-convergence has, however, lead to a suggestion that fragmentation may occur for very long cooling times. Meru & Bate (2012) do, however, extrapolate their results to suggest that they may be tending towards convergence and that the critical value may be around $\beta \sim 30$.

One issue with the idea that fragmentation could occur for very long cooling times is, however, based on our basic understanding of the underlying processes. As discussed above, in the absence of other heating sources, there is a relationship between cooling time and α, such that as the cooling time gets longer, the effective α value gets smaller. However, this also implies that the perturbation amplitudes get smaller. So, one issue would be how do discs fragment if the perturbations are tiny?

Also, if fragmentation at longer cooling times requires higher resolution, then it implies that this fragmentation is occurring on scales that are not resolved by the lower resolution simulations. However, the Jeans mass/radius at $Q \sim 1$ is typically resolved in at least part of the simulation domains for these lower resolution simulations (Rice et al. 2014; Tsukamoto et al. 2015). This would suggest that at these longer cooling times, large amplitude, small-scale perturbations are driving $Q \ll 1$ and hence producing fragments that are unresolved in the lower resolution simulations. In shearing sheet simulations, however, the power typically decreasing with increasing wavenumber, k, as shown in Figure 7. The issue then becomes why at long cooling times fragmentation happens at scales where there is much less power than at the scales at which fragmentation occurs when the cooling time is short?

Consequently, Rice, Forgan, & Armitage (2012), suggest that the effect is largely numerical and is related to the manner in which cooling has typically been implemented in SPH simulations. More recently, an SPH simulation with 10 million particles and with a modified form of the standard β cooling formalism, was shown not to fragment for $\beta = 10$ (Rice et al. 2014), well below the values suggested by Meru & Bate (2011) and Meru & Bate (2012). Michael et al. (2012) also considered convergence in a three-dimensional grid-based code and, although they didn't test for the fragmentation boundary specifically, they did find that simulations with $\alpha < 0.06$ typically did not fragment. Hence, it seems that it's important to be careful of how the cooling is implemented and to avoid implementations that may promote fragmentation.

There has been some attempt to try and see if there is a physical explanation for this lack of convergence. In particular, is there evidence that simulations are suppressing fragmentation at small scales? One possibility, suggested by Young & Clarke (2015), is that there are two modes of fragmentation. One mode operates at the free-fall time and requires rapid cooling. The other is a mode in which fragments form via quasi-static collapse and requires that these fragments then contract to a size where they can survive disruption by spiral shock waves. Young & Clarke (2015) suggest that many two-dimensional simulations smooth gravity on the Jeans scale, inhibiting contraction to smaller scales, and hence preventing this second mode from operating. However, they also show that even if this second mode does operate, it probably still requires relatively fast cooling ($\beta \lesssim 12$), but with quite a large uncertainty.

Part of the motivation behind Young & Clarke (2015) was a suggestion that turbulent discs are never stable against fragmentation (Hopkins & Christiansen 2013) and, in particular, that rare high-amplitude, small-scale fluctuations could lead to gravitational collapse. This suggestion, however, is not entirely consistent with what we're discussing here. The Hopkins & Christiansen (2013) analysis considered discs that were typically assumed to be isothermal, essentially implying that $\beta = 0$.

Also, the turbulence they were considering was not the gravito-turbulence driven by the instability itself. The standard analysis of self-gravitating discs considers a scenario in which the quasi-turbulent structures are a self-consistent response to the growth of the instability and depend—as discussed already—on the thermal balance in the disc. Hopkins & Christiansen (2013) were considering a situation in which isothermal turbulence is present in a disc in which self-gravity might also operate. In fact, the presence of such turbulence—if the disc were non-isothermal—should act as an extra heating source, which would then be expected to reduce the magnitude of the gravitational instability, rather than enhance it (Rice et al. 2011).

3.3. Triggered fragmentation

Even though the analysis in Hopkins & Christiansen (2013) is not quite consistent with the basic picture presented here, it is still interesting from the perspective of whether or not fragmentation could be triggered. This could be via some other form of turbulence (as suggested by Hopkins & Christiansen 2013) or via some kind of external perturbation. Some early work (e.g., Boffin et al. 1998; Watkins et al. 1998) suggested that stellar encounters could promote fragmentation. Such early simulations typically, however, assumed an isothermal equation of state, which is known to promote fragmentation (Pickett et al. 1998, 2000). Later work suggested that the inclusion of compressive and shock heating (Nelson et al. 2000; Lodato et al. 2007) would tend to prohibit encounter-driven fragmentation. This was further confirmed using simulations with a realistic radiative transfer formalism (Forgan et al. 2009).

The basic argument is that, typically, the encounter will compress the disc on a dynamical timescale, which is much shorter than the thermal timescale. Consequently, the encounter would be expected to heat—and stablise—the disc, rather than produce fragmentation. Fragmentation would only be expected to occur if the cooling time can be reduced by the compression (Whitworth et al. 2007). This may be possible if the temperature change pushes the gas into a regime where the opacity is reduced (the opacity-gap). However, the corresponding increase in gas density tends to act in the opposite sense, and so this may still make triggered fragmentation unlikely (Johnson & Gammie 2003). There has, however, been a recent suggestion (Meru 2015) that fragmentation in the inner parts of discs may be triggered by fragments that form, initially, in the outer parts of discs, and Thies et al. (2010) suggest that tidal encounters might trigger fragmentation in massive, extended protostellar discs. Hence, there may be scenarios under which fragmentation could be triggered.

3.4. Stochasticity

Paardekooper (2012) have suggested that fragmentation may be stochastic. Their basic argument is that a quasi-steady gravito-turbulent state consists of weak shocks and transient clumps that will contract on the cooling timescale (β). If such a clump can avoid being disrupted by one of the spiral shocks, then it could survive to become a bound fragment. For example, Figure 8—taken from Paardekooper (2012)—shows the evolution of the maximum surface density ($\Sigma_{\max}/\Sigma_o$) and α in a shearing sheet simulation in which the disc appears to initially settle into a quasi-steady, self-regulated state. This is illustrated by both the maximum density and α settling to reasonably constant values, as indicated by the dashed line in the lower panel. However, at $\Omega t \sim 300$, α starts to decrease and the maximum density increases to a value a few hundred times greater than the initial value, Σ_o. This indicates that this system has undergone fragmentation after first appearing to settle to a quasi-steady state, and may be indicative of stochasticity.

Paardekooper (2012) find that transient clumps are present in all of their simulations (up to $\beta = 50$) but that fragmentation only actually happens up to $\beta = 20$. Hopkins & Christiansen (2013) suggest that, given sufficient time, stochastic fragmentation may actually occur both if the cooling time is very long and the Q value is very high. However, as we suggest above, the scenario they're considering isn't quite the same as that being discussed here.

Young & Clarke (2016) argue that this stochasticity is consistent with the idea behind their suggestion that there are two modes of fragmentation (Young & Clarke 2015). In discs with rapid cooling, fragmentation can occur on the free-fall time. In discs that are cooling slowly (for example, $\beta > 10$–20),

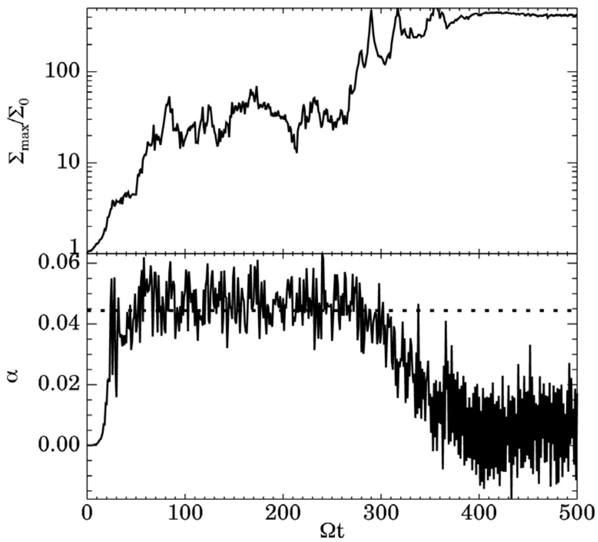

Figure 8. Figure—from Paardekooper (2012)—showing the maximum surface density relative to the initial surface density (Σ_{max}/Σ_o—top panel) and α (bottom panel) against time, from a shearing sheet simulation that initially appears to have settled into a quasi-steady, self-regulated state. However, at $\Omega t \sim 300$, α starts to decrease and the maximum surface density starts increasing, eventually reaching values many hundreds of times greater than it was initially. The system, which appeared to have settled into a quasi-steady state, has now undergone fragmentation and this may be indicative of stochasticity.

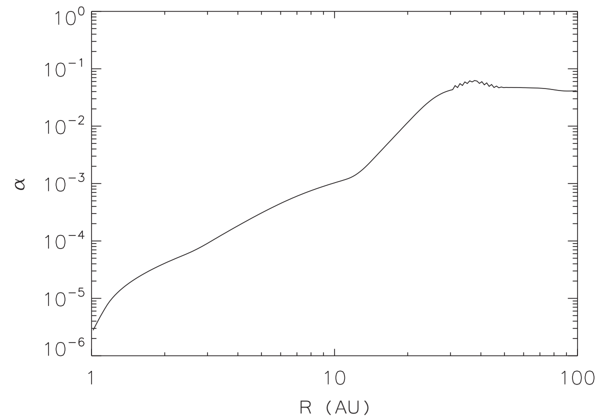

Figure 9. Figure showing α against radius in a quasi-steady, self-gravitating disc with realistic cooling. The inner disc is optically thick, leading to long cooling times and very small α values. The α value increases with radius, reaching value where fragmentation might be possible in the outer parts of the disc. Hence, we expect fragmentation to only be possible in the outer parts of protostellar accretion discs.

fragmentation can still occur if a clump can cool and contract prior to encountering a spiral shock. Young & Clarke (2016) find that the wait time between shocks is not strongly dependent on the cooling time. Given that clumps contract on the cooling timescale, this means that the likelihood of fragmentation decreases with increasing cooling time. They also find an exponential decay in the likelihood of a patch remaining unshocked, and so the probability of a transient clump contracting sufficiently so as to survive a shock encounter is effectively zero for very long cooling times. Given this, even if fragmentation is stochastic, this is unlikely to substantially change the conditions under which fragmentation actually occurs. Similarly, Baehr & Klahr (2015) impose a cooling implementation that attempts to incorporate optical depth effects, and find that this also appears to inhibit stochastic fragmentation.

3.5. Does fragmentation actually occur?

If we focus specifically on discs around young stars (protostellar discs), then we can determine where fragmentation could happen, and even if it does actually happen. It is likely that the inner regions of protostellar discs are optically thick and, consequently, have long cooling times. This is illustrated in Figure 9 which shows α plotted against radius for a quasi-steady disc in which the cooling time at each radius is calculated using the actual optical depth, rather than by simply assuming some β-value *a priori*. The α values are extremely small in the inner disc but increase with increasing radius, reaching values where fragmentation may be possible in the outer parts of the disc. Consequently, we expect fragmentation to only occur in the outer parts of such systems (Matzner & Levin 2005; Rafikov 2005; Stamatellos & Whitworth 2008), beyond ∼30–50 AU.

Even if fragmentation is stochastic, or can occur for longer cooling times than we currently think, this has little impact when it comes to fragmentation in protostellar discs. The radial dependence of the cooling time is very strong in the inner parts of these discs (Clarke 2009; Rice & Armitage 2009). Even a factor of a few in the cooling time for fragmentation has little effect on where fragmentation likely happens (Young & Clarke 2016). If the opacity is significantly lower than interstellar, then the fragmentation radius could to move towards the inner parts of the disc (Meru & Bate 2010). However, it is still unlikely that fragmentation can take place inside ∼10–20 AU in such discs (Meru & Bate 2010).

In addition to where we might expect fragmentation to occur, we can also estimate typical fragment masses (Boley et al. 2010; Forgan & Rice 2011). Typically, the initial fragment masses are a few Jupiter masses. Given that we expect fragmentation to only occur in the outer parts of protostellar discs, and to have initial masses of at least a few Jupiter masses, it's been suggested this could explain some of the directly imaged planets on wide orbits (Kratter et al. 2010; Nero & Bjorkman 2009). It's also been suggested that in some cases such objects could rapidly spiral into the inner disc (Baruteau, Meru, & Paardekooper 2011; Michael, Durisen, & Boley 2011; Malik et al. 2015) and could lose mass via tidal stripping (Nayakshin 2010; Boley et al. 2010). This could potentially then be the origin of some of the known exoplanets, even terrestrial/rocky exoplanets if the cores can survive disruption (Nayakshin 2011). If not, the disruption of an embryo could lead to the formation of planetesimal belts (Nayakshin & Cha 2012) or change the overall disc chemistry (Boley et al. 2010). In some cases, however, it may be possible for the protoplanet to open a gap and survive (Stamatellos 2015).

Forgan & Rice (2013), however, argue that typically fragments are either destroyed, or remain at large radii with relatively high masses (many Jupiter masses). According to their analysis, the formation of planetesimal belts or terrestrial planets via this mechanism is also rare. The high destruction rate, however, is consistent with other simulations (Boley & Durisen 2010; Zhu et al. 2012) and could lead to outbursts, such as the FU Orionis phenomenon (Vorobyov & Basu 2005; Dunham & Vorobyov 2012). Galvagni & Mayer (2014), however, suggest that a large fraction of the clumps could survive inward migration and that such a process could explain a reasonable fraction of the known population of 'hot' Jupiters and other gas giant planets.

In addition to the rapid inspiral of fragments, it's also possible that objects forming via fragmentation at large radii could be scattered by stellar, or other, companions (Forgan, Parker, & Rice 2015). Some of these could then be circularised onto a short-period orbit, via tidal interactions with the parent star (Lin, Bodenheimer, & Richardson 1996), producing either a 'hot' Jupiter or a proto-hot Jupiter (Dawson & Johnson 2012). Rice et al. (2015), however, argue that the population of known 'hot' and proto-hot Jupiters is inconsistent with this being common and suggests that disc fragmentation rarely forms planetary-mass objects. This is also consistent with the relatively low frequency of known wide orbit planetary and brown dwarf companions (Biller et al. 2013; Brandt et al. 2014).

4 DISCUSSION AND CONCLUSION

We've tried here to summarise our current understanding of the evolution of self-gravitating accretion discs, focusing—in particular—on those around young's protostars. When self-gravity is important, we expect such discs to either settle into a quasi-steady, self-regulated state in which heating and cooling are in balance and $Q \sim 1$ (Paczynski 1978), or—if the instability is particularly strong—to fragment to form bound objects.

Typically, the boundary between a disc fragmenting, or settling into a quasi-steady state, is determined by the rate at which it is losing energy, with rapid cooling potentially leading to fragmentation (Gammie 2001). The exact boundary does, however, depend on whether or not there is an additional heating source (Rice et al. 2011), and on the mass of the disc relative to the mass of the central object; due to the global nature of the gravitational instability (Balbus & Papaloizou 1999), very massive discs may avoid fragmentation under conditions that may lead to fragmentation in less massive discs (Lodato & Rice 2005; Forgan et al. 2011). Similarly, very massive discs may not settle into a true quasi-steady, self-regulated state, but instead show variability, and are only quasi-steady in a time averaged sense. It is also possible that fragmentation itself is stochastic and could, at times, occur for longer cooling times than this basic analysis suggests (Paardekooper 2012).

In protostellar discs, however, the exact fragmentation boundary is not that relevant when it comes to determining where fragmentation might happen. The inner parts of such discs are so optically thick, and cool so slowly, that fragmentation is only actually likely in the outer parts of these discs (Rafikov 2005). Even if fragmentation could occur for slightly longer cooling times than originally thought, or if fragmentation is stochastic, this is unlikely to significantly influence where fragmentation can occur (Young & Clarke 2016). Although, reductions in opacity can bring the fragmentation region closer to the central star (Meru & Bate 2010; Rogers & Wadsley 2012), it's still unlikely close to the central star. Essentially, it is very difficult for there to be conditions such that fragmentation could occur inside $\sim$20 AU.

There are, however, still aspects that are uncertain. It might seem that the lack of convergence in self-gravitating disc simulations (Meru & Bate 2011) is most likely numerical (Lodato & Clarke 2011; Meru & Bate 2012; Rice et al. 2012, 2014), but this is still not fully resolved. Similarly, there is evidence for stochasticity (Paardekooper 2012) and a potential explanation for this (Young & Clarke 2016), but there are still aspects of this that are uncertain. However, our basic understanding seems robust. Self-gravitating discs will tend to settle into a quasi-steady, self-regulated state in which heating balances cooling, in which the transport of angular momentum can be described as pseudo-viscous (Lodato & Rice 2004), and in which the amplitude of the perturbations depends on the magnitude of the effective viscosity α (Cossins et al. 2009). If the perturbations become very large, then the disc may undergo fragmentation.

It's also unclear as to whether or not fragmentation ever actually occurs in protostellar discs. There are suggestions (Kratter et al. 2010; Nero & Bjorkman 2009) that it may explain some of the directly imaged exoplanets on wide-orbits (Marois et al. 2008) and suggestions that some may spiral inwards to smaller orbital radii than where they form (Nayakshin 2010). On the other hand, the paucity of directly imaged planets on wide-orbits (Biller et al. 2013) and the properties of the known closer-in exoplanets, suggests that disc fragmentation may rarely form planetary-mass objects (Rice et al. 2015).

However, disc self-gravity may still play a crucial role in angular momentum transport during the earliest stages of star formation, and the resulting density waves may then play a role in the growth of planet building material (Rice et al. 2004; Gibbons, Rice, & Mamatsashvili 2012). So, even if the gravitational instability does not play a significant role in forming planets directly, it may still play an important role in the growth of the building blocks crucial for planet formation.

ACKNOWLEDGEMENTS

The author would like to thank the organisers of the Disc Dynamics and Planet Formation meeting, held in Cyprus, that partly motivated this paper. The author would also like to thank the reviewer for a very thorough and detailed review that helped to improve the paper, and would like to acknowledge useful comments from Phil Armitage. KR gratefully acknowledges support from STFC grant ST/M001229/1. Some of the research leading to these results also received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 313014 (ETAEARTH).

REFERENCES

Baehr, H., & Klahr, H. 2015, ApJ, 814, 155
Balbus, S. A., & Papaloizoum, J. C. B. 1999, ApJ, 521, 650
Baruteau, C., Meru, F., & Paardekooper, S.-J. 2011, MNRAS, 416, 1971
Bell, K. R., & Lin, D. N. C. 1994, ApJ, 427, 987
Biller, B. A., et al. 2013, ApJ, 777, 160
Boffin, H. M. J., Watkins, S. J., Bhattal, A. S., Francis, N., & Whitworth, A. P. 1998, MNRAS, 300, 1189
Boley, A. C., & Durisen, R. H. 2010, ApJ, 724, 618
Boley, A. C., Hayfield, T., Mayer, L., & Durisen, R. H. 2010, Icarus, 207, 509
Boley, A. C., Mejía, A. C., Durisen, R. H., Cai, K., Pickett, M. K., & D'Alessio, P. 2006, ApJ, 651, 517
Bonnell, I. A., & Rice, W. K. M. 2008, Science, 321, 1060
Boss, A. P. 1998, Nature, 393, 141
Boss, A. P. 2000, ApJ, 536, L101
Boss, A. P. 2001, ApJ, 551, L167
Boss, A. P. 2003, ApJ, 599, 577
Brandt, T. D., et al. 2014, ApJ, 794, 159
Cai, K., Durisen, R. H., Michael, S., Boley, A. C., Mejía, A. C., Pickett, M. K., & D'Alessio, P. 2006, ApJ, 636, L149
Clarke, C. J. 2009, MNRAS, 396, 1066
Cossins, P., Lodato, G., & Clarke, C. J. 2009, MNRAS, 393, 1157
Dawson, R. I., & Johnson, J. A. 2012, ApJ, 756, 122
Dong, R., Hall, C., Rice, K., & Chiang, E. 2015, ApJ, 812, L32
Dunham, M. M., & Vorobyov, E. I. 2012, ApJ, 747, 52
Forgan, D., & Rice, K. 2011, MNRAS, 417, 1928
Forgan, D., Rice, K., Cossins, P., & Lodato, G. 2011, MNRAS, 410, 994
Forgan, D., Rice, K., Stamatellos, D., & Whitworth, A. 2009, MNRAS, 394, 882
Forgan, D. H., Parker, R. J., & Rice, K. 2015, MNRAS, 447, 836
Forgan, D. H., & Rice, K. 2013, MNRAS, 432, 3168
Gammie, C. F. 2001, ApJ, 553, 174
Galvagni, M., & Mayer, L. 2014, MNRAS, 437, 2909
Gibbons, P. G., Rice, W. K. M., & Mamatsashvili, G. R. 2012, MNRAS, 426, 1444
Haisch, K. E., Lada, E. A., & Lada, C. J. 2001, ApJ, 553, L153
Hopkins, P. F., & Christiansen, J. L. 2013, ApJ, 776, 48
Johnson, B. M., & Gammie, C. F. 2003, ApJ, 597, 131
Kratter, K. M., Murray-Clay, R. A., & Youdin, A. N. 2010, ApJ, 710, 1375
Laughlin, G., & Bodenheimer, P. 1994, ApJ, 436, 335
Laughlin, G., & Rozcyczka, M. 1996, ApJ, 456, 279
Levin, Y., & Beloborodov, A. M. 2003, ApJ, 590, L33
Lin, D. N. C., Bodenheimer, P., & Richardson, D. C. 1996, Nature, 380, 606
Lin, D. N. C., & Pringle, J. E. 1987, MNRAS, 225, 607
Lodato, G., & Clarke, C. J. 2011, MNRAS, 413, 2735
Lodato, G., Meru, F., Clarke, C. J., & Rice, W. K. M. 2007, MNRAS, 374, 590
Lodato, G., & Rice, W. K. M. 2004, MNRAS, 351, 630
Lodato, G., & Rice, W. K. M. 2005, MNRAS, 358, 1489
Malik, M., Meru, M., Mayer, L., & Meyer, M. 2015, ApJ, 802, 56
Marois, C., Macintosh, B., Barman, T., Zuckerman, B., Song, I., Patience, J., Lafreniére, D., & Doyon, R. 2008, Science, 322, 1348
Matzner, C. D., & Levin, Y. 2005, ApJ, 628, 817
Mayer, L., Lufkin, G., Quinn, T., & Wadsley, J. 2007, ApJ, 661, L77
Mejía, A. C., Durisen, R. H., Pickett, M. K., & Cai, K. 2005, ApJ, 619, 1098
Meru, F. 2015, MNRAS, 454, 2529
Meru, F., & Bate, M. R. 2010, MNRAS, 406, 2279
Meru, F., & Bate, M. R. 2011, MNRAS, 411, L1
Meru, F., & Bate, M. R. 2012, MNRAS, 427, 2022
Michael, S., Steiman-Cameron, T. Y., Durisen, R. H., & Boley, A. C. 2012, ApJ, 746, 98
Michael, S., Durisen, R. H., & Boley, A. C. 2011, ApJ, 737, L42
Murray, J. R. 1996, MNRAS, 279, 402
Nayakshin, S. 2010, MNRAS, 408, L36
Nayakshin, S. 2011, MNRAS, 410, L1
Nayakshin, S., & Cha, S.-H. 2012, MNRAS, 423, 2104
Nayakshin, S., Cuadra, J., & Springel, V. 2007, MNRAS, 379, 21
Nelson, A. F., Benz, W., & Ruzmaikina, T. V. 2000, ApJ, 529, 357
Nero, D., & Bjorkman, J. E. 2009, ApJ, 702, L163
Paardekooper, S.-J. 2012, MNRAS, 421, 3286
Paardekooper, S.-J., Baruteau, C., & Meru, F. 2011, MNRAS, 416, L65
Paczynski, B. 1978, AcA, 28, 91
Pickett, B. K., Benz, W., Adams, F. C., & Arnett, D. 1998, ApJ, 502, 342
Pickett, B. K., Cassen, P., Durisen, R. H., & Link, R. 2000, ApJ, 529, 1034
Pickett, B. K., Mejía, A. C., Durisen, R. H., Cassen, P. M., Berry, D. K., & Link, R. P. 2003, ApJ, 590, 1060
Pringle, J. E. 1981, ARA&A, 19, 137
Rafikov, R. R. 2005, ApJ, 621, 69
Rafikov, R. R. 2009, ApJ, 704, 281
Rice, K., Lopez, E., Forgan, D., & Biller, B. 2015, MNRAS, 454, 1940
Rice, W. K. M., & Armitage, P. J. 2009, MNRAS, 396, 2228
Rice, W. K. M., Armitage, P. J., Bate, M. R., & Bonnell, I. A. 2003, MNRAS, 339, 1025
Rice, W. K. M., Armitage, P. J., Mamatsashvili, G. R., Lodato, G., & Clarke, C. J. 2011, MNRAS, 418, 1356
Rice, W. K. M., Forgan, D. H., & Armitage, P. J. 2012, MNRAS, 420, 1640
Rice, W. K. M., Lodato, G., & Armitage, P. J. 2005, MNRAS, 364, 56
Rice, W. K. M., Lodato, G., Pringle, J. E., Armitage, P. J., & Bonnell, I. A. 2004, MNRAS, 355, 543
Rice, W. K. M., Mayo, J. H., & Armitage, P. J. 2010, MNRAS, 402, 1740

Rice, W. K. M., Paardekooper, S.-J., Forgan, D. H., & Armitage, P. J. 2014, MNRAS, 438, 1593
Rogers, P. D., & Wadsley, J. 2012, MNRAS, 423, 1896
Safronov, V. S. 1960, Ann. Astrophys., 23, 979
Shakura, N. I., & Sunyaev, R. A. 1973, A&A, 24, 337
Stamatellos, D. 2015, ApJ, 810, L11
Stamatellos, D., & Whitworth, A. P. 2008, MNRAS, 392, 413
Stamatellos, D., Whitworth, A. P., Bisbas, T., & Goodwin, S. 2007, A&A, 475, 37
Thies, I., Kroupa, P., Goodwin, S. P., Stamatellos, D., & Whitworth, A. P. 2010, ApJ, 717, 577
Toomre, A. 1964, ApJ, 139, 1217
Tsukamoto, Y., Takahashi, S. Z., Machida, M. N., & Inutsuka, S. 2015, MNRAS, 446, 1175
Vorobyov, E. I., & Basu, S. 2005, ApJ, 633, L137
Watkins, S. J., Bhattal, A. S., Boffin, H. M. J., Francis, N., & Whitworth, A. P. 1998, MNRAS, 300, 1205
Whitworth, A. P., Bate, M. R., Nordlund, Å., Reipurth, B., & Zinnecker, H. 2007, in Protostars and Planets V, ed. B. Reipurth, & K. Keil (Tucson: University of Arizona Press), 459s
Young, M. D., & Clarke, C. J. 2015, MNRAS, 451, 3987
Young, M. D., & Clarke, C. J. 2016, MNRAS, 455, 1438
Zhu, Z., Hartmann, L., Nelson, R. P., & Gammie, C. F. 2012, ApJ, 746, 110