Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **494,** 624–641 (2020)
Advance Access publication 2020 March 09

doi:10.1093/mnras/staa659

How do bound star clusters form?

Mark R. Krumholz [1,2,3,4]★ and Christopher F. McKee[5]

[1]*Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2611, Australia*
[2]*ARC Centre of Excellence for Astronomy in Three Dimensions (ASTRO-3D), Canberra, ACT 2611, Australia*
[3]*Institut für Theoretische Astrophysik, Zentrum für Astronomie, Universität Heidelberg, D-69120 Heidelberg, Germany*
[4]*Max Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany*
[5]*Departments of Physics and Astronomy, University of California, Berkeley, CA 94720, USA*

Accepted 2020 March 3. Received 2020 March 3; in original form 2019 August 26

ABSTRACT

Gravitationally bound clusters that survive gas removal represent an unusual mode of star formation in the Milky Way and similar spiral galaxies. While forming, they can be distinguished observationally from unbound star formation by their high densities, virialized velocity structures, and star formation histories that accelerate towards the present, but extend multiple free-fall times into the past. In this paper, we examine several proposed scenarios for how such structures might form and evolve, and carry out a Bayesian analysis to test these models against observed distributions of protostellar age, counts of young stellar objects relative to gas, and the overall star formation rate of the Milky Way. We show that models in which the acceleration of star formation is due either to a large-scale collapse or a time-dependent increase in star formation efficiency are unable to satisfy the combined set of observational constraints. In contrast, models in which clusters form in a 'conveyor belt' mode where gas accretion and star formation occur simultaneously, but the star formation rate per free-fall time is low, can match the observations.

Key words: stars: formation – ISM: kinematics and dynamics – open clusters and associations: general – galaxies: star clusters: general.

1 INTRODUCTION

The typical outcome of star formation in spiral galaxies is not a gravitationally bound star cluster. In the Milky Way, Lada & Lada (2003) were among the first to point out that the number of observed star clusters at ages from 10 to 100 Myr is a factor of ∼10 smaller than one would expect if every observed gas-embedded star-forming clump were to go on to become a cluster of comparable mass. The natural explanation for this discrepancy is that most of the young stars that we observe in star-forming regions are in fact unbound, or will become so once the gas is removed, and that we count them as cluster members at young ages simply because they have not yet had time to drift apart. Extensive surveys of external galaxies echo this conclusion, with counts of star clusters as a function of age implying that no more than 5–10 per cent of stars that form will remain part of a gravitationally bound structure several tens of Myr after formation (e.g. Adamo et al. 2015; Johnson et al. 2016; Chandar et al. 2017; Messa et al. 2018; see the recent review by Krumholz, McKee & Bland-Hawthorn 2019, for additional references).

Thus regions of star formation that do go on to become gravitationally bound clusters must be special in some way. Recent observational advances offer significant hints about how such regions might be special. Regions that go on to become bound clusters do not appear to represent a distinct class of gas cloud, such that most clouds unbind entirely and a small minority remain mostly bound. Instead, many star-forming regions appear to consist of a dense inner part that contains a minority of the mass, which is likely to go on to become bound, and an extended outer part whose stars will drift apart. The inner regions that go on to become bound are distinguishable in several ways.

First, they appear to feature extended star formation histories. Low-density star-forming regions that are ∼10 pc in size or larger tend to have stellar populations whose ages are comparable to their crossing times (Elmegreen 2000; Kruijssen et al. 2019), suggesting a relatively rapid formation process. By contrast, the densest regions of star formation, with sizes ∼1 pc, have star formation histories that are significantly more extended compared to their dynamical times. The best-studied example is the Orion Nebula Cluster (ONC), where the free-fall time in the central 1 pc is ≈0.6 Myr (Da Rio, Tan & Jaehnig 2014), but there is extensive evidence that star formation has been ongoing for a significantly longer period (e.g. Reggiani et al. 2011; Jaehnig, Da Rio & Tan 2015; Da Rio et al. 2016; Beccari et al. 2017). Star formation in this region appears to be accelerating (Palla & Stahler 2000; Huff & Stahler

★ E-mail: mark.krumholz@gmail.com

2006; Caldwell & Chang 2018), but even accounting for this effect most stars are significantly older than a free-fall time – using the kinematically-selected sample and estimated ages of Kounkel et al. (2018, 2019) find that 50 per cent of the stars in the ONC are older than 3 free-fall times, and 10 per cent are older than 10 free-fall times. However, the ONC appears to be typical in this regard: similarly extended but accelerating star formation histories have been observed in NGC 6530 (more than 25 per cent of stars older than three free-fall times Prisinzano et al. 2019), Perseus (Azimlu, Martínez-Galarza & Muench 2015), Taurus, and ρ Ophiuchus (Caldwell & Chang 2018), though the last three of these regions are still highly gas-dominated, and it is therefore unclear if they will in fact reach star formation efficiencies sufficient to produce a bound cluster.

Secondly, the regions with extended star formation histories are also distinct kinematically. While most young stars still embedded in their parent molecular clouds are characterised by unrelaxed density and velocity distributions (e.g. Fűrész et al. 2008; Tobin et al. 2009), the density distribution in the central 1 pc of the ONC can be fit reasonably well by an isothermal, spherically symmetric King (1962) model (Hillenbrand & Hartmann 1998), and the velocity distribution is virialised (Kim et al. 2019). This region is neither expanding or contracting, and there is no evidence for a population of stars on primarily-radial orbits that are plausibly falling toward or escaping from it (Ward & Kruijssen 2018; Kuhn et al. 2019).

While regions like the ONC appear to be distinct in some respects, they also share one very significant commonality with the more extended envelopes around them. The density of young stellar objects (YSOs) increases smoothly with gas surface density, with no clear breaks at the densities or radii that correspond to the shift from unrelaxed, fractal stellar distributions to relaxed, virialized ones (Gutermuth et al. 2011). Once one normalizes the gas surface density by the free-fall time, it correlates remarkably tightly with YSO count; there is a near-linear relationship between YSO mass and gas mass normalized by free-fall time with a scatter of only $\approx$0.3–0.4 dex across orders of magnitude in mass and density (Krumholz, Dekel & McKee 2012; Lada et al. 2013; Evans, Heiderman & Vutisalchavakul 2014; Heyer et al. 2016; Ochsendorf et al. 2017 – see fig. 10 of Krumholz et al. 2019 for a compilation of results). One can interpret this correlation as describing the efficiency of star formation: the star formation efficiency (SFE) per free-fall time is $\epsilon_{\rm ff} = \dot{M}_*/(M_{\rm g}/t_{\rm ff})$, where $M_{\rm g}$ and $t_{\rm ff}$ are the gas mass and free-fall time. If there are $N_{\rm YSO}$ YSOs associated with this gas that have a mean mass $M_{\rm YSO}$ and that remain spectrally identifiable as such for a time $t_{\rm YSO}$, then the star formation rate (SFR) must be $\dot{M}_* \approx N_{\rm YSO} M_{\rm YSO}/t_{\rm YSO}$. All published studies based on YSO counts give $\epsilon_{\rm ff} \approx 0.01$, with $\lesssim$0.4 dex scatter; the low value of $\epsilon_{\rm ff}$ and the extended star formation histories in regions that become bound are likely related, since a low $\epsilon_{\rm ff}$ region is likely to become bound only if it forms stars long enough to reach a respectable total SFE, and for the stars formed to dynamically relax (Kruijssen 2012). In contrast, ratios of far-infrared or free–free luminosity to gas mass give a much larger dispersion (Lee, Miville-Deschênes & Murray 2016; Vutisalchavakul, Evans & Heyer 2016; Ochsendorf et al. 2017). However, these results depend critically upon the procedure used to match regions of FIR or free–free emission to spatially separated molecular clouds, with differing matching procedures yielding results that differ by up to $\sim$1 dex (Krumholz et al. 2019). Given the consistency of the much more direct YSO results, we regard them as more reliable.

Since regions like the ONC appear to be distinct from other star-forming regions in some ways but not others, and appear to evolve distinctly from the bulk of the young stellar population once star formation ends and gas is cleared, it is interesting to attempt to characterise the star formation process in these regions. Our goal in this paper is to examine a variety of proposed scenarios for star cluster formation that may be found in the literature, construct simple mathematical descriptions for them, and confront them with the wide variety of observational results that we have just outlined. We present the models to which we are interested in comparing, and outline a general framework for describing them, in Section 2. In Section 3, we compare these models to the observations outlined above, determining where they succeed and where they fail. We summarize our findings in Section 4.

2 FRAMEWORK FOR CLUSTER FORMATION

We now sketch out some simple, general models for how star clusters might form. Before beginning this exercise, it is important to understand that our goal is not to examine fully self-consistent and detailed models for star cluster formation. Even purely analytic or semi-analytic models for cluster formation and cloud evolution (e.g. Goldbaum et al. 2011; Zamora-Avilés, Vázquez-Semadeni & Colín 2012; Zamora-Avilés & Vázquez-Semadeni 2014; Lee et al. 2016; Lee & Hennebelle 2016b) generally include complex prescriptions for the time evolution of cloud mass, density, velocity dispersion, star formation activity, the effects of stellar feedback, and similar details. Comparing observations to such models is in general very difficult, because the models have many moving parts and contain numerous tuneable parameters. Our goal instead is to develop cartoons that capture some of the main qualitative features of models that have been proposed in the literature, but that are analytically computable and have relatively few free parameters, so that we can carry out statistical comparisons to observation. This means that we will simply prescribe the evolution of parameters such as cloud mass and density, rather than trying to compute them fully self-consistently. As we introduce the individual models below, we will point out features of the more complex published models they are intended to capture.

All the softwares used to produce all the plots and analysis found in this paper are publicly available at https://bitbucket.org/krumholz/km19/.

2.1 General framework

We begin by characterizing a gas cloud that is in the process of forming a star cluster in terms of its instantaneous gas mass $M_{\rm g}$ and mean density ρ; it is convenient to characterize the latter in terms of the free-fall time $t_{\rm ff} = \sqrt{3\pi/32G\rho}$. Both $M_{\rm g}$ and ρ can in general be functions of time. At any instant, the cloud forms stars at a rate

$$\dot{M}_* = \epsilon_{\rm ff} \frac{M_{\rm g}}{t_{\rm ff}}. \quad (1)$$

For simplicity we will generally only worry about mean quantities, but we note that, if instead of a uniform cloud one considers a cloud where the density profile is a power law $\rho \propto r^{-k_\rho}$, and one assumes that equation (1) holds locally (i.e. at every point the star formation density obeys $\dot{\rho}_* = \epsilon_{\rm ff}\rho/t_{\rm ff}$), then the sole modification to equation (1) is that $\epsilon_{\rm ff}$ is increased by a factor of $[2/(2-k_\rho)][(3-k_\rho)/3]^{3/2}$, which is of the order of unity unless k_ρ is very close to 2.

In addition to star formation, the cloud can gain mass by accretion and lose it by ejection of mass by stellar feedback. We take the mass removal rate by feedback to be proportional to the SFR $\dot{M}_{\rm fb} = \eta \dot{M}_*$, while the accretion rate $\dot{M}_{\rm acc}$ is an input parameter; here η is the

usual mass loading factor.[1] The total mass of gas and stars therefore evolve following:

$$\dot{M}_{\rm g} = \dot{M}_{\rm acc} - (1+\eta)\,\epsilon_{\rm ff}\frac{M_{\rm g}}{t_{\rm ff}}, \qquad \dot{M}_* = \epsilon_{\rm ff}\frac{M_{\rm g}}{t_{\rm ff}}. \tag{2}$$

In principle both $\dot{M}_{\rm acc}$ and η can, like $t_{\rm ff}$, be a function of time.

2.2 Scenarios of star formation

Having established this general framework, we now consider a range of scenarios for how a star cluster might be assembled. We plot example histories for each model in Fig. 1, and summarize the models and their key free parameters in Table 1.

2.2.1 Static cloud (ST)

Our first scenario is the simplest possible, a non-accreting cloud with constant $t_{\rm ff}$ that begins with an initial gas mass $M_{\rm g} = M_{\rm g,0}$ and starts forming stars at time $t = 0$. We refer to this as the static, or ST model, hereafter. Of course, if the density and free-fall time are constant, but the gas mass is not, then this means that the cloud is not static in terms of its radius; since the data to which we will compare below do not include detailed information on the spatial structure of stellar populations, however, the constant free-fall time is the property that is relevant for our purposes. Physically, this would correspond to a situation where cloud assembly is rapid compared to the process of star formation, or where a cloud is assembled in a state where it cannot form stars immediately. As first pointed out by Ginsburg et al. (2012) and Longmore et al. (2014, also see Walker et al. 2016; Urquhart et al. 2018), such a scenario can be ruled out for at least the most massive young clusters observed in the Milky Way, on the grounds that there are no observed gas clouds dense and massive enough to be the progenitors of the most massive clusters. On the other hand, Krumholz et al. (2019) point out that there is no such difficulty for clusters near the Galactic Centre, and in this region there do indeed appear to be very massive and dense molecular clouds with little or no star formation activity such as 'the Brick' (Longmore et al. 2013; Rathborne et al. 2014). These have been hypothesized to remain quiescent until star formation within them is triggered by a pericentre passage around Sgr A* (Kruijssen, Dale & Longmore 2015), and thus they represent potential exemplars of the static cloud scenario, though recent observations of infall in at least some of these objects suggest something closer to one of the alternative scenarios we describe below (Barnes et al. 2019).

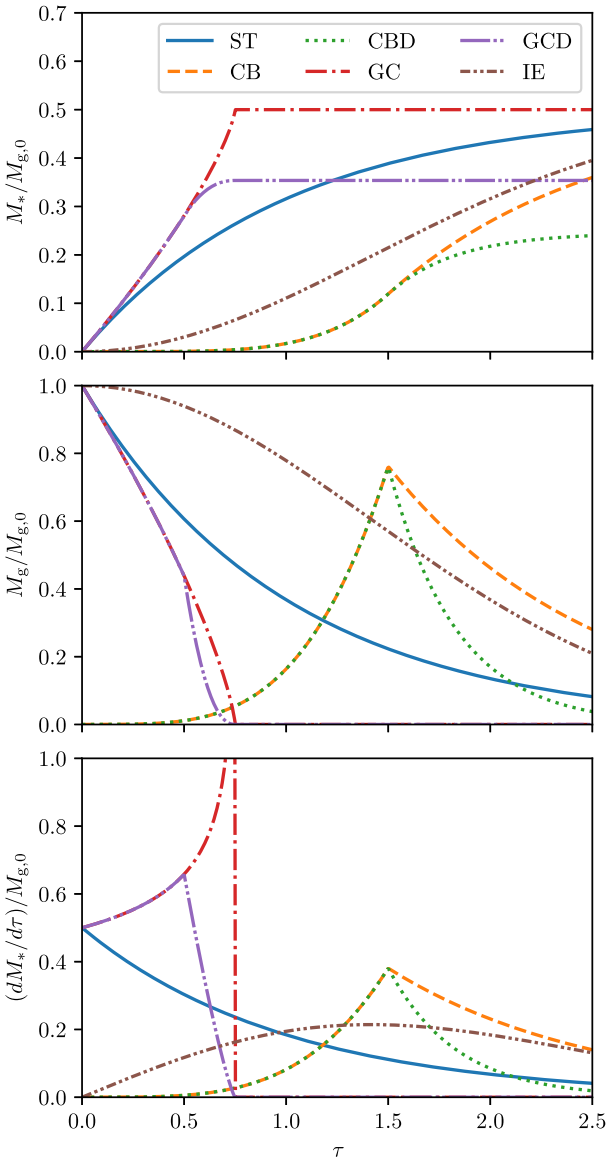

Figure 1. Example evolutionary histories of stellar mass (top), gas mass (middle), and SFR (bottom) for each of the models discussed in the paper (as indicated in the legend). For the purposes of this plot, we use $\eta = 1$ in all models. For CB and CBD, we use $\tau_{\rm acc} = 1.5$ and $p = 3$, for CBD and GCD we use $\eta_{\rm d} = 5$, for GC we use $\xi = 1$ and $\tau_{\rm coll} = 0.75$, for GCD we use $\xi = 1$, $\tau_{\rm coll} = 0.75$, and $\tau_{\rm fb} = 0.5$, and for IE we use $\chi = 0.5$ and $\delta = 1$. See main text for definitions of the various parameters.

Since there is no mass accretion in this model, $\dot{M}_{\rm acc} = 0$, and we will also assume η is constant, the solution to equation (2) is trivial:

$$M_* = \frac{M_{\rm g,0}}{1+\eta}\left(1 - {\rm e}^{-\tau}\right) \qquad M_{\rm g} = M_{\rm g,0}{\rm e}^{-\tau}, \tag{3}$$

where $\tau = t/t_{\rm sf}$ and

$$t_{\rm sf} = \frac{t_{\rm ff}}{(1+\eta)\epsilon_{\rm ff}} \tag{4}$$

is the star formation time-scale; this is the natural time-scale over which the star formation process occurs, and the cloud is converted to stars or dispersed. The final SFE, defined as the ratio of final

[1]Our choice to parametrize mass-loss in terms of a mass-loading factor η, so that the mass removal rate is proportional to the star formation rate, differs from some other simple models (e.g. Lee et al. 2016) in which the mass removal rate is taken to be proportional to the total stellar mass. As discussed in Dekel & Krumholz (2013), which of these approximations is preferable depends on how the duration of star formation compares to the duration of the feedback mechanisms that dominate mass removal – $\dot{M}_{\rm fb} \propto \dot{M}_*$ is preferable if star formation is extended compared to feedback, $\dot{M}_{\rm fb} \propto M_*$ if not. The dominant feedback mechanisms in a forming star cluster are likely to be protostellar outflows (on for $\approx$0.1 Myr) for clusters that do not contain O stars, and photoionization or radiation pressure (on for $\approx$3 Myr) for those that do (Krumholz et al. 2019). Below we will compare to data on two star clusters, NGC 6530 and the ONC. In NGC 6530, the duration of star formation is $\approx$1–2 Myr, and there are no O stars; in the ONC, there is an O star, but the duration of star formation is $\approx$3–4 Myr. Since both of these systems have star formation durations comparable to or longer than the corresponding feedback duration, we prefer to model the mass removal rate as proportional to the instantaneous star formation rate.

Table 1. Summary of models and their parameters. Note that not all of these parameters are independent, and in cases where parameters are related, we list the relationship in the table.

Model name	Abbreviation	Parameter	Meaning
Parameters common to all models		$\epsilon_{\rm ff}$	Star formation efficiency per free-fall time
		η	Mass loading factor
		$t_{\rm ff}$	Free-fall time
		$t_{\rm sf}$	Star formation time-scale, $t_{\rm sf} = t_{\rm ff}/[(1+\eta)\epsilon_{\rm ff}]$
Static cloud	ST	–	
Conveyor belt	CB	p	Accretion rate versus time $\dot{M}_{\rm acc} \propto t^p$
		$t_{\rm acc}$	Duration of accretion flow; dimensionless time $\tau_{\rm acc} \equiv t_{\rm acc}/t_{\rm sf}$
Conveyor belt + dispersal	CBD	p	Accretion rate versus time $\dot{M}_{\rm acc} \propto t^p$
		$t_{\rm acc}$	Duration of accretion flow; dimensionless time $\tau_{\rm acc} \equiv t_{\rm acc}/t_{\rm sf}$
		$\phi_{\rm d}$	Ratio of $1+\eta$ during dispersal phase to value during accretion phase
Global collapse	GC	$t_{\rm coll}$	Collapse time; dimensionless time $\tau_{\rm coll} \equiv t_{\rm coll}/t_{\rm sf}$
		$t_{\rm ff,\,0}$	Free-fall time at onset of star formation; for this model $t_{\rm sf} \equiv t_{\rm ff,\,0}/[(1+\eta)\epsilon_{\rm ff}]$
		ξ	Ratio of collapse time-scale to free-fall time-scale, $t_{\rm coll} = 2t_{\rm ff,\,0}/\xi$
Global collapse + dispersal	GCD	$t_{\rm coll}$	Collapse time; dimensionless time $\tau_{\rm coll} \equiv t_{\rm coll}/t_{\rm sf}$
		$t_{\rm ff,0}$	Free-fall time at onset of star formation; for this model $t_{\rm sf} \equiv t_{\rm ff,0}/[(1+\eta)\epsilon_{\rm ff}]$
		ξ	Ratio of collapse time-scale to free-fall time-scale, $t_{\rm coll} = 2t_{\rm ff,0}/\xi$
		$t_{\rm fb}$	Time at which feedback increases; dimensionless $\tau_{\rm fb} \equiv t_{\rm fb}/t_{\rm sf}$
		$\phi_{\rm d}$	Ratio of $1+\eta$ during dispersal phase to value during earlier phase
Increasing efficiency	IE	δ	Efficiency per free-fall time varies as $\epsilon_{\rm ff} = \epsilon_{\rm ff,0}(t/t_{\rm ff})^\delta$
		$\epsilon_{\rm ff,0}$	Value of $\epsilon_{\rm ff}$ at $t = t_{\rm ff}$; for this model, $t_{\rm sf} \equiv t_{\rm ff}/[(1+\eta)\epsilon_{\rm ff,0}]$
		χ	Ratio of star formation time-scale to free-fall time-scale, $\chi = t_{\rm sf}/t_{\rm ff}$

stellar mass to total mass of gas available for star formation, is

$$\epsilon_* \equiv \frac{M_*}{M_{\rm g,0}} = \frac{1}{1+\eta}. \tag{5}$$

2.2.2 *Conveyor belt (CB)*

The absence of gas clouds as massive and dense as the densest star clusters, as noted in Section 2.2.1, led Longmore et al. (2014) to propose a 'conveyor belt' model where gas accretion occurs simultaneously with cluster formation, so that the full mass of the gas cloud is never assembled at a single time; observations that regions such as the ONC frequently sit at the intersections of filaments supports this picture (Motte, Bontemps & Louvet 2018). In this picture, stars may form in both the filaments and in the central hub, but stars that wind up as part of a bound cluster at the end of the star formation process are mostly those that form in the central hub. This hub region is continually re-supplied by accretion of gas from the filaments. For the purposes of this paper, and for the data sets to which we will compare below, we are primarily interested in what happens in the hub.

In principle the region fed by a conveyor belt could be static, expanding, or contracting. Numerical simulations and analytic calculations by a number of authors (e.g. Klessen & Burkert 2000; Goldbaum et al. 2011; Matzner & Jumper 2015; Lee & Hennebelle 2016a, b) suggest that, as long as the accretion rate is high enough that a cloud's growth time is comparable to its free-fall time, the inflow supplies enough energy for the density and SFR per free-fall time to remain roughly constant for multiple free-fall times. Examples of such models include Goldbaum et al. (2011, their fig. 3), Zamora-Avilés & Vázquez-Semadeni (2014, the 10^5 and 10^6 $M_\odot$ models shown in their fig. 1), and Lee & Hennebelle (2016b, their fig. 6): in all of these models, the free-fall time varies by no more than a factor of $\sim$2 over multiple cloud free-fall times. For this reason we will assume constant $t_{\rm ff}$ and $\epsilon_{\rm ff}$. We refer to this model as conveyor belt, or CB, hereafter.

We abstract this model as having an initial gas mass of zero, and an accretion rate that varies in time as a power law t^p. We generically expect $p > 0$, since gravity-driven accretion rates generally rise with time until the reservoir of mass is exhausted; Goldbaum et al. (2011) show that pressureless collapse of a reservoir of constant surface density that becomes gravitationally unstable naturally produces $p \approx 3$; Lee & Hennebelle (2016b) find a similar value of p while protoclusters are small compared to their parent reservoirs, but that this tapers to $p \approx 0$ once $\gtrsim 10$ per cent of the parent reservoir has been accreted. We will adopt the Goldbaum et al. value of $p = 3$ as our fiducial choice, but for completeness we give the model result for general p, by taking the accretion rate to be

$$\dot{M}_{\rm acc} = H(t_{\rm acc} - t)(p+1)\frac{M_{\rm g,0}}{t_{\rm acc}}\left(\frac{t}{t_{\rm acc}}\right)^p, \tag{6}$$

where $M_{\rm g,0}$ is the total mass that will eventually reach the protocluster, $t_{\rm acc}$ is the time over which accretion happens, and $H(x)$ is the Heaviside step function. The initial conditions are $M_{\rm g} = M_* = 0$. With this accretion rate, equation (2) has the following analytic solutions for any non-negative integer p:

$$M_* = \begin{cases} \dfrac{M_{\rm g,0}}{(1+\eta)(p+2)\tau_{\rm acc}^{p+1}} g(\tau,\, p+2), & \tau \le \tau_{\rm acc} \\ M_*(\tau_{\rm acc}) + \dfrac{M_{\rm g}(\tau_{\rm acc})}{1+\eta}\left(1 - {\rm e}^{-\tau+\tau_{\rm acc}}\right), & \tau > \tau_{\rm acc} \end{cases} \tag{7}$$

$$M_g = \begin{cases} \dfrac{M_{\rm g,0}}{\tau_{\rm acc}^{p+1}} g(\tau,\, p+1), & \tau \le \tau_{\rm acc} \\ M_{\rm g}(\tau_{\rm acc}){\rm e}^{-\tau+\tau_{\rm acc}}, & \tau > \tau_{\rm acc} \end{cases}, \tag{8}$$

where for $p \ge 1$

$$g(\tau, p) = p{\rm e}^{-\tau}\int_0^\tau \tau'^{(p-1)}{\rm e}^{\tau'}\,{\rm d}\tau', \tag{9}$$

$$= p!\left[(-1)^p{\rm e}^{-\tau} - \sum_{i=1}^{p}\frac{(-1)^i}{(p-i)!}\tau^{p-i}\right]. \tag{10}$$

Here $\tau = t/t_{\rm sf}$ as in Section 2.2.1, $\tau_{\rm acc} = t_{\rm acc}/t_{\rm sf}$, and we made use of the relations

$$\int_0^\tau g(\tau', p)d\tau' = \frac{g(\tau, p+1)}{p+1}, \tag{11}$$

$$= \tau^p - g(\tau, p). \tag{12}$$

To get a feeling for the magnitude of $g(\tau, p)$, we note that $g(\tau, 1) = 1 - \mathrm{e}^{-\tau}$ and that $g(\tau, 2) = 2(\tau - 1 + \mathrm{e}^{-\tau})$. The approximation

$$g(\tau, p) \simeq \frac{\tau^p}{1 + \tau/p} \tag{13}$$

is accurate to better than 15 per cent. Next, observe that equation (11) implies

$$\frac{\mathrm{d}g(\tau, p+1)}{\mathrm{d}\tau} = (p+1)g(\tau, p). \tag{14}$$

In turn, this relation implies that $g(\tau, p)$ is a monotonically increasing function of τ since equation (9) implies that $g(\tau, p)$ is positive. It follows from equation (8) that the gas mass increases monotonically until the accretion stops.

At times $\tau \gg \tau_{\rm acc}$, the SFE in the conveyor belt model approaches $\epsilon_* = 1/(1+\eta)$, exactly as in the static cloud case, but the star formation history is different. This model satisfies the observational constraint that originally motivated it, in that the gas mass need never be large compared to the final stellar mass. Indeed, the final stellar mass (achieved in the limit $\tau \to \infty$) is $M_{*,\rm f} = M_{\rm g,0}/(1+\eta)$ regardless of the accretion history, while the maximum gas mass (achieved when $\tau = \tau_{\rm acc}$) is $M_{\rm g,max} \approx M_{\rm g,0}/[1 + \tau_{\rm acc}/(p+1)]$. Thus as long as $\tau_{\rm acc} \gtrsim \eta$, the maximum gas mass will be comparable to or smaller than the final stellar mass.

An important feature of this conveyor belt model is that *star formation always accelerates while gas is accreting*, provided $p \geq 0$. With the aid of equation (14), we find that the acceleration in the stellar mass is

$$\ddot{M}_* = \frac{p+1}{(1+\eta)\tau_{\rm acc}^{p+1}} \left(\frac{M_{\rm g,0}}{t_{\rm sf}^2}\right) g(\tau, p+2) \qquad \tau \leq \tau_{\rm acc}, \tag{15}$$

which is always positive, as noted above. Such acceleration appears to be demanded by the observations (Palla & Stahler 2000).

2.2.3 Conveyor belt plus rapid dispersal (CBD)

A slight variation on the standard conveyor belt model is to note that, as pointed out by Goldbaum et al. (2011), mass-loss rates are likely sensitive to the strength of the confining ram pressure from accretion. Consequently, it makes sense to adopt a mass loading factor that increases significantly once accretion ceases, leading to more rapid dispersal. We refer to conveyor belt models in which dispersal after the end of accretion is rapid as conveyor belt plus dispersal (CBD) models hereafter. From the standpoint of our simple analytic models, we can model this by setting η to one value during the accretion phase, $t < t_{\rm acc}$, and to another value $\eta_{\rm d} > \eta$ during the dispersal phase, $t > t_{\rm acc}$. In all other respects this model is identical to the simple conveyor belt model of Section 2.2.2. The solution to equation (2) in this case is modified only slightly from that given by equations (7) and (8):

$$M_* = \begin{cases} \dfrac{M_{\rm g,0}}{(1+\eta)(p+2)\tau_{\rm acc}^{p+1}} g(\tau, p+2), & \tau \leq \tau_{\rm acc} \\ M_*(\tau_{\rm acc}) + \dfrac{M_{\rm g}(\tau_{\rm acc})}{1+\eta_{\rm d}} \left(1 - e^{-\phi_{\rm d}(\tau - \tau_{\rm acc})}\right) & \tau > \tau_{\rm acc} \end{cases} \tag{16}$$

$$M_g = \begin{cases} \dfrac{M_{\rm g,0}}{\tau_{\rm acc}^{p+1}} g(\tau, p+1), & \tau \leq \tau_{\rm acc} \\ M_{\rm g}(\tau_{\rm acc})\mathrm{e}^{-\phi_{\rm d}(\tau - \tau_{\rm acc})} & \tau > \tau_{\rm acc} \end{cases}, \tag{17}$$

where

$$\phi_{\rm d} \equiv \frac{1+\eta_{\rm d}}{1+\eta} \tag{18}$$

can be thought of as representing the ratio of star formation efficiencies during and after the accretion phase. This model shares the key feature of the conveyor belt model: there is no need to assemble a cloud as massive as the final star cluster all at once, since the histories are identical up to the end of the accretion phase, but then a smaller fraction of the remaining gas mass is converted to stars than in the standard conveyor belt case. To be precise, the final SFE is

$$\epsilon_* = \frac{1}{1+\eta}\left[1 - \left(\frac{\phi_{\rm d} - 1}{\phi_{\rm d}}\right)\frac{g(\tau_{\rm acc}, p+1)}{\tau_{\rm acc}^{p+1}}\right]. \tag{19}$$

Equations (11) and (12) imply that the ratio $g(\tau, p+1)/\tau^{p+1}$ is strictly smaller than unity for any $\tau > 0$ since $g(\tau, p+2) > 0$, so the final SFE is between $1/(1+\eta)$ and $1/(1+\eta_{\rm d})$.

2.2.4 Global collapse (GC)

The observation that star formation accelerates could be a reflection of gas accumulation, as in the CB or CBD models, but it could also be a result of the star formation process itself. An example of such a model is the global collapse (GC) scenario proposed by a number of authors (e.g. Zamora-Avilés & Vázquez-Semadeni 2014; Kuznetsova, Hartmann & Ballesteros-Paredes 2015, 2018; Vázquez-Semadeni, González-Samaniego & Colín 2017; Vázquez-Semadeni et al. 2019). The central idea of GC models is that clouds are assembled in a low-density state but then undergo a global collapse. Consequently, the mean free-fall time, rather than remaining constant, systematically decreases on a free-fall time-scale as the mean density rises. The combination of an apparently extended star formation history and an accelerating SFR is then taken to be due to the decreasing free-fall time: stars that form at early times may have ages comparable to the free-fall time of the system when the formed, but this can be significantly longer than the free-fall time of the system at the time when it is observed. Moreover, as the system gets denser, the free-fall time decreases and thus star formation accelerates.

In terms of the hub-and-filament geometry frequently observed in star-forming regions, and discussed in Section 2.2.2, the difference between the CB (or CBD) and GC models is the assumed time evolution of the hubs. In the CB model, the hub is assumed to remain at roughly constant density over many free-fall times, so that any acceleration of star formation is due to the mass of the hub increasing, not due to its density rising. By contrast, in GC the hub is assumed to be in a process of collapse on a dynamical time-scale (even if it is also accreting), so that the density rises with time, and this accounts for most or all of the increase SFR with time. Examples of published models in the latter category include the 10^3 or 10^4 $\mathrm{M}_\odot$ cases shown in fig. 1 of Zamora-Avilés & Vázquez-Semadeni (2014), where, once the clouds grow massive enough, the density runs away to infinity on roughly a free-fall time-scale.

Mathematically we can represent this model by assuming that the mean density obeys

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \xi \frac{\rho}{t_{\rm ff}(\rho)}, \tag{20}$$

where $t_{\rm ff}(\rho) = \sqrt{3\pi/32G\rho}$ is the free-fall time at the current density. The constant ξ specifies how fast the cloud contracts compared to the free-fall time-scale, with higher ξ corresponding

to more rapid contraction. The value of ξ will depend at least partly on geometry – $\xi \approx 1$ is expected for 3D structures, but values as small as $\sim$0.1 are possible for highly flattened geometries if one interprets ρ as the density internal to the structure (Toalá, Vázquez-Semadeni & Gómez 2012); however, note that if one interprets ρ as the mean density of a spherical structure of the same size, as is frequently done when interpreting observations, then $\xi \gtrsim 1$ even for flattened structures. For a cloud that starts at density ρ_0 at time $t = 0$, the density and free-fall time evolve as

$$\rho = \frac{\rho_0}{(1-x)^2} \qquad t_{\rm ff} = t_{\rm ff,0}\,(1-x), \tag{21}$$

where $x = t/t_{\rm coll}$, $t_{\rm coll} = 2t_{\rm ff,0}/\xi$ is the time at which the cloud reaches infinite density, and $t_{\rm ff,0} = \sqrt{3\pi/32G\rho_0}$ is the initial free-fall time.

Inserting this non-constant free-fall time into equation (2), holding η and $\epsilon_{\rm ff}$ constant, and solving subject to the initial condition that $M_{\rm g} = M_{\rm g,\,0}$ and $M_* = 0$ at $t = 0$, we obtain

$$M_* = \frac{M_{\rm g,0}}{1+\eta}\begin{cases} 1-(1-x)^{\tau_{\rm coll}}, & x < 1 \\ 1, & x \geq 1 \end{cases} \tag{22}$$

$$M_{\rm g} = M_{\rm g,0}\begin{cases} (1-x)^{\tau_{\rm coll}}, & x < 1 \\ 0, & x \geq 1 \end{cases}. \tag{23}$$

The quantity

$$\tau_{\rm coll} = \frac{2(1+\eta)\epsilon_{\rm ff}}{\xi} \tag{24}$$

is the dimensionless time at which the cloud collapses to infinite density and $t_{\rm ff} \to 0$, where we have non-dimensionalized time using $\tau = t/t_{\rm sf}$ as before, but we now define $t_{\rm sf} = t_{\rm ff,0}/[(1+\eta)\epsilon_{\rm ff}]$ (cf. equation 4), i.e. we define $t_{\rm sf}$ using the initial free-fall time since $t_{\rm ff}$ is non-constant. Half the stars have formed and half the gas has been consumed at a time

$$t_{1/2} = \left(1 - \frac{1}{2^{1/\tau_{\rm coll}}}\right) t_{\rm coll}, \tag{25}$$

and correspondingly the free-fall time then is

$$t_{\rm ff,\,1/2} = \frac{t_{\rm ff,0}}{2^{1/\tau_{\rm coll}}}. \tag{26}$$

For $\tau_{\rm coll} \gtrsim 1$, half the stars form at a rate not that different from the initial rate. Indeed, in the limit $\xi \ll 1$, and thus $\tau_{\rm coll} \gg 1$, the GC model approaches the ST model, since the collapse then becomes slow compared to star formation. (Conversely, in the limit $\epsilon_{\rm ff} \to 1$, the ST and CB models become qualitatively similar to GC, since then all gas is converted to stars on a dynamical time-scale.) More generally, the rate at which the SFR changes is

$$\ddot{M}_* = \frac{M_{\rm g,0}}{(1+\eta)t_{\rm sf}^2}\left(\frac{1-\tau_{\rm coll}}{\tau_{\rm coll}}\right)\left(1-\frac{\tau}{\tau_{\rm coll}}\right)^{\tau_{\rm coll}-2}, \tag{27}$$

so star formation accelerates with time ($\ddot{M}_* > 0$) only if $\tau_{\rm coll} < 1$. The final SFE is $\epsilon_* = 1/(1+\eta)$, exactly as in the ST or CB models.

2.2.5 *Global collapse plus dispersal (GCD)*

Just as the CBD model adds a more rapid dispersal phase (i.e. a larger value of η) to CB, one can similarly posit a GC model with rapid dispersal at its end. In the CBD model the natural cause of an increase is the removal of confinement by the accretion flow. In GC there is no similar natural breakpoint, but a number of authors (e.g. Vázquez-Semadeni et al. 2019, and references therein) have posited that the stellar initial mass function (IMF) is time-dependent, so that massive stars only form late in the collapse process. If this hypothesis were correct, it would naturally cause the mass loading factor to increase at later times. Mathematically, we model this by introducing two new free parameters: $\phi_{\rm d}$, which is defined exactly as for the CBD model (equation 18) as the ratio of star formation efficiencies before and after massive star feedback 'turns on', and $t_{\rm fb}$, which represents the time at which this happens.

If we let η be the mass loading parameter prior to $t < t_{\rm fb}$, $\eta_{\rm d} = \phi_{\rm d}(1+\eta) - 1$ be the mass loading factor from $t_{\rm fb} < t < t_{\rm coll}$, and continue to use equation (21) to describe the evolution of the free-fall time, the solution to equation (2) is

$$M_* = \frac{M_{\rm g,0}}{1+\eta}\cdot \begin{cases} 1-(1-x)^{\tau_{\rm coll}}, & x < x_{\rm fb} \\ \phi_{\rm d}^{-1}\,(1-x_{\rm fb})^{\tau_{\rm coll}}\left[1-\left(\frac{1-x}{1-x_{\rm fb}}\right)^{\phi_{\rm d}\tau_{\rm coll}}\right] + & \\ \qquad 1-(1-x_{\rm fb})^{\tau_{\rm coll}}, & x_{\rm fb} \leq x < 1 \\ 1-\left(\frac{\phi_{\rm d}-1}{\phi_{\rm d}}\right)(1-x_{\rm fb})^{\tau_{\rm coll}}, & x \geq 1 \end{cases} \tag{28}$$

$$M_{\rm g} = M_{\rm g,0}\cdot \begin{cases} (1-x)^{\tau_{\rm coll}}, & x < x_{\rm fb} \\ (1-x_{\rm fb})^{\tau_{\rm coll}}\left(\frac{1-x}{1-x_{\rm fb}}\right)^{\phi_{\rm d}\tau_{\rm coll}}, & x_{\rm fb} \leq x < 1 \\ 0, & x \geq 1 \end{cases}, \tag{29}$$

where $x_{\rm fb} = t_{\rm fb}/t_{\rm coll}$. The final SFE is

$$\epsilon_* = \frac{1}{1+\eta}\left[1-\left(\frac{\phi_{\rm d}-1}{\phi_{\rm d}}\right)(1-x_{\rm fb})^{\tau_{\rm coll}}\right]. \tag{30}$$

As with CBD (cf. equation 19), the factor inside the square brackets is strictly negative, and thus the final SFE is lower than in the corresponding model without the disruption phase. Star formation continues accelerating during the gas clearing phase only if $\phi_{\rm d}\tau_{\rm coll} < 1$; otherwise it decelerates.

2.2.6 *Increasing star formation efficiency (IE)*

A final potential mechanism to explain why star formation accelerates in protoclusters like the ONC is to posit that this is an intrinsic part of the star formation process itself. Lee, Chang & Murray (2015) and Murray & Chang (2015) argue that, rather than being constant, $\epsilon_{\rm ff}$ increases with time in star-forming regions as $\epsilon_{\rm ff} \propto t^{\delta}$, with $\delta \approx 1$; we refer to this as the increasing efficiency (IE) model. Although somewhat similar to the GC model, the two are conceptually distinct in that star formation accelerates in the GC model because the mean density rises with time, while in the IE model it accelerates even though the mean density remains constant because the star formation process itself becomes more efficient. Mathematically, the two models differ in their predicted rate of acceleration. Caldwell & Chang (2018) argue that the IE model provides a good fit to observed star formation histories in resolved clusters, and Lee et al. (2016) and Ochsendorf et al. (2017) argue it provides a good fit to the observed ratio of ionizing luminosity to CO luminosity, though, as we note above, the quality of the agreement is extremely sensitive to the choice of procedure for matching up non-co-spatial molecular gas and H II regions.

For the purposes of comparing this model to data, we adopt the same parametrization as Lee et al. (2016): $\epsilon_{\rm ff} = \epsilon_{\rm ff,\,0}(t/t_{\rm ff})^{\delta}$. Thus $\epsilon_{\rm ff,0}$ represents the value of $\epsilon_{\rm ff}$ one free-fall time after the onset of star formation. While the theoretical models of Lee et al. (2015) and Murray & Chang (2015) give $\delta = 1$, we will allow δ to be a free parameter from 0 to 3 when we fit to observations below. The solution to equation (2) for arbitrary $\delta \geq 0$, holding η and $t_{\rm ff}$

constant, subject to the initial conditions $M_g = M_{g,0}$ and $M_* = 0$ at $t = 0$, is

$$M_* = \frac{M_{g,0}}{1+\eta}\left[1 - \exp\left(-\frac{\chi^\delta \tau^{1+\delta}}{1+\delta}\right)\right] \quad (31)$$

$$M_g = M_{g,0}\exp\left(-\frac{\chi^\delta \tau^{1+\delta}}{1+\delta}\right), \quad (32)$$

where $\tau = t/t_{\rm sf}$, $t_{\rm sf} = t_{\rm ff}/[(1+\eta)\epsilon_{\rm ff,0}]$ (i.e. we define $t_{\rm sf}$ using the value of $\epsilon_{\rm ff}$ at 1 free-fall time; c.f. equation 4), and $\chi = t_{\rm sf}/t_{\rm ff} = 1/[(1+\eta)\epsilon_{\rm ff,0}]$. The final SFE is $\epsilon_* = 1/(1+\eta)$, exactly as in the static model. The average efficiency with which stars form is

$$\begin{aligned}\bar{\epsilon}_{\rm ff} &= \frac{1+\eta}{M_{g,0}}\int_0^\infty \epsilon_{\rm ff}(\tau)\dot{M}_*(\tau)\,{\rm d}\tau \\ &= [\chi(1+\delta)]^{\delta/(1+\delta)}\,\Gamma\left(1+\frac{\delta}{1+\delta}\right)\epsilon_{\rm ff,0}. \end{aligned} \quad (33)$$

For typical parameters in this model, $\delta = 1$ and $\chi = 50$, this gives $\bar{\epsilon}_{\rm ff} \approx 8.9\epsilon_{\rm ff,0}$, so most stars form at an efficiency substantially higher than that which prevails for the first free-fall time. Intuitively, this makes sense: in this model there are a relatively long period of near-quiescence when $\epsilon_{\rm ff}$ is small and few stars form, but this is followed by a burst of activity after $\epsilon_{\rm ff}$ becomes large; most stars form during this final burst. Quantitatively, the second derivative of the SFR is

$$\ddot{M}_* = \frac{M_{g,0}}{(1+\eta)t_{\rm sf}^2}\left[\chi^\delta \tau^{\delta-1}\left(\delta - \chi^\delta \tau^{\delta+1}\right)\exp\left(-\frac{\chi^\delta \tau^{1+\delta}}{1+\delta}\right)\right]. \quad (34)$$

The sign of $\ddot{M}_*$ therefore depends on $\delta - \chi^\delta \tau^{\delta+1}$; for sufficiently small τ this term is positive, and star formation accelerates. Later on, as gas is depleted, this term becomes negative and star formation decelerates.

3 CONFRONTATION WITH OBSERVATIONS

Having outlined the various models, we now compare them to observations.

3.1 Star formation histories

3.1.1 Data set

The first observation to which we are interested in comparing is the observed distribution of stellar ages in young clusters; as discussed in Section 1, working through the implications of the observed extended but accelerating star formation histories in such regions is one of our primary motivations in this work. For our observational data set, we select two young open clusters: the ONC and NGC 6530. We focus on these two because they both offer very clean, high-quality data: membership lists determined from *Gaia* 6D phase-space data plus other ancillary indicators, and ages determined from spectroscopy, with star-by-star extinction corrections. The free-fall time in the ONC is $t_{\rm ff,ONC} \approx 0.6$ Myr as determined from dynamical modelling by Da Rio et al. (2014). For NGC 6530, Prisinzano et al. (2019) measure a stellar velocity dispersion of $\sigma_{\rm NGC\,6530} = 2.42\ {\rm km\ s^{-1}}$, and the effective radius of the cluster is 0.1° (Kharchenko et al. 2013), which translates to 2.3 pc for the best-fitting distance of 1.32 kpc obtained by Prisinzano et al.. Thus the crossing time is $t_{\rm cr} = r_{\rm NGC\,6530}/\sigma_{\rm NGC\,6530} = 0.93$ Myr. For a virialized object, the free-fall time is approximately half the crossing time (Tan, Krumholz & McKee 2006), so we adopt $t_{\rm ff,NGC6530} = 0.5$ Myr.

For our stellar ages in NGC 6530, we use the fits provided by Prisinzano et al. (2019). For the ONC, we must select down from the full catalogue of Kounkel et al. (2018), since their study covers the entire Orion star-forming region and includes multiple populations across a large volume. For this study, we select stars from their catalog that are within 1 pc in projection of θ^1 C (the same radius within which we have estimated the free-fall time), and that are kinematically identified as part of the Orion A population. We take the ages of these stars from Kounkel et al., using only the ages based on spectroscopic determinations, since those based on colour are unreliable in the ONC due to high extinction. After applying these cuts, our sample consists of 185 stars in the ONC and 395 stars in NGC 6530.

In addition to the age estimates themselves, in order to carry out a meaningful statistical analysis we must have some understanding of the uncertainties in the measurements. Uncertainties in the ages of young stars has been a topic of considerable debate in the literature in recent years, and we refer to the readers to the reviews by Soderblom et al. (2014), Jeffries (2017), and Krumholz et al. (2019) for a detailed discussion. Young stellar ages are always subject to a systematic uncertainty of ∼0.1–0.3 dex in the *absolute* age scale coming from the choice of pre-main-sequence tracks. However, there is significantly less uncertainty in the *relative* ages of stars (e.g. Reggiani et al. 2011), which is the quantity of concern for us, since we are interested in the star formation history – a shift in absolute age just amounts to a rescaling of the time-scales.

Relative age uncertainties come from a variety of factors, depending on the age-dating method. Uncertainties larger than ≈0.2–0.3 dex can be ruled out by independent methods of constraining dispersions of stellar age (e.g. radii derived from rotation or gravity-sensitive spectral features – Jeffries 2007; Da Rio et al. 2016; Prisinzano et al. 2019), but a range of estimates below this limit have been published (e.g. Preibisch 2012; Da Rio et al. 2016; Prisinzano et al. 2019). For this work we adopt the results of Prisinzano et al. (2019): we take the error in log age to be a Gaussian with a width $\sigma = 0.13$ dex and a bias $b = -0.05$ dex (i.e. true stellar ages are on average 0.05 dex older than estimated ones). The systematic bias is due to unresolved binarity, which increases luminosity at fixed effective temperature, and thus tends to bias age estimates low. We have experimented with other choices of these parameters, subject to the overall constraint that the total error cannot exceed ≈0.2−0.3 dex, and we find that the posterior PDFs for some parameters can be sensitive to the exact choice of σ and b, as are quantitative measures of relative goodness of fit such as the Akaike information criterion. Since we do not understand the true error distribution in detail, we will for this reason limit our analysis to general features that are robust against plausible changes in σ or b.

3.1.2 Likelihood function

We wish to compare the observed age distribution to that predicted by our various candidate models. To this end, we now compute a likelihood function, which gives the probability density of the data given the model. For convenience we summarise the meanings of various parameters that we introduce in this calculation in Table 2.

Table 2. Definitions of parameters used in computing the stellar age distribution likelihood function.

Parameter	Meaning
$t_{\rm clust}$	Age of cluster (time since onset of star formation)
t_*	True age of a star
$t_{*,\rm obs}$	Observationally estimated stellar age (including errors)
σ	Dispersion of stellar age error distribution
b	Bias in the stellar age error distribution
$t_{\rm ff,\,clust}$	Present-day free-fall time in cluster
$\epsilon_{*,\,\rm clust}$	Present-day star formation efficiency, $M_*(t_{\rm clust})/M_{\rm g,0}$
$f_{\rm g,\,clust}$	Present-day gas fraction, $M_{\rm g}/(M_{\rm g}+M_*)$ at $t=t_{\rm clust}$

For a cluster formation model with stellar mass as a function of dimensionless time, $M_*(\tau)$, the distribution of log stellar ages that will be seen a time when the cluster age is $t_{\rm clust}$ (i.e. a time $t_{\rm clust}$ after the onset of star formation) is

$$\frac{{\rm d}p}{{\rm d}\log t_*} = (\ln 10)\,\frac{t_* M_*'(\tau_{\rm clust}-\tau_*)}{t_{\rm sf} M_*(\tau_{\rm clust})} = (\ln 10)\,\epsilon_{\rm ff}\left(\frac{t_*}{t_{\rm ff}}\right)\frac{M_g(\tau_{\rm clust}-\tau_*)}{M_*(\tau_{\rm clust})}, \quad (35)$$

where t_* is the stellar age, $\tau_{\rm clust}=t_{\rm clust}/t_{\rm sf}$, and $\tau_*=t_*/t_{\rm sf}$ are the dimensionless cluster and stellar ages, respectively, and $M_*'={\rm d}M_*/{\rm d}\tau$. The factor of ln 10 is to ensure that the PDF is properly normalized to have unit integral over all $\log t_*$. The stellar mass versus dimensionless time, $M_*(\tau)$, is given by equations (3), (7), (16), (22), (28), and (31), for the ST, CB, CBD, GC, GCD, and IE models, respectively; the corresponding gas masses, $M_g(\tau)$, are given by equations (3), (8), (17), (23), (29), and (32).

Note that, in the GC and GCD models, $t_{\rm ff}$ is also a function of $\tau_{\rm clust}-\tau_*$ (equation 21). These models produce a double-peaked profile in the distribution ${\rm d}p/{\rm dlog}\,t_*$; equation (35) shows that the age distribution is proportional to $\tau_* M_*'(\tau_{\rm clust}-\tau_*)$, or, in terms of the parameter τ in Fig. 1, $(\tau_{\rm clust}-\tau)M_*'(\tau)$. Reference to Fig. 1 shows that this leads to a double peak in the GC and GCD models, with one peak at $\tau\sim\tau_{\rm clust}$ and a second at $\tau\sim\tau_{\rm clust}-\tau_{\rm coll}$ or $\tau\sim\tau_{\rm clust}-\tau_{\rm fb}$.

To incorporate the effects of errors, we convolve the true age distribution with the error distribution. Following our discussion in Section 3.1.1, we parametrize the uncertainty distribution in log age as a biased Gaussian, i.e. for a star whose *true* log age is $\log t_*$, the distribution of *measured* log ages $\log t_{*,\rm obs}$ is

$$f(\log t_{*,\rm obs}\mid \log t_*) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(\log t_* - \log t_{*,\rm obs}+b\right)^2}{2\sigma^2}\right], \quad (36)$$

where b is the bias and σ is the dispersion, and both b and σ are in units of dex. The full distribution of observed ages is therefore given by

$$\frac{{\rm d}p}{{\rm d}\log t_{*,\rm obs}} = \int_{-\infty}^{\infty}\left(\frac{{\rm d}p}{{\rm d}\log t_*}\right) f(\log t_{*,\rm obs}\mid \log t_*)\,{\rm d}\log t_*. \quad (37)$$

We evaluate this integral numerically via Fourier transform, since it is equivalent to the convolution of the true stellar age distribution ${\rm d}p/{\rm dlog}\,t_*$ with a Gaussian. The log likelihood function $\mathcal{L}$ is simply the probability density of the data given the model:

$$\log\mathcal{L} = \sum_{i=1}^{N}\log\left(\frac{{\rm d}p_*}{{\rm d}\log t_{*,\rm obs}}\right)_{t_{*,\rm obs}=t_i}, \quad (38)$$

where t_i is the age estimated for the ith star in the observed sample.

Our stellar age distributions as written depend on two-dimensional quantities: the cluster age $t_{\rm clust}$, and the star formation time-scale $t_{\rm sf}$ that scales between physical times t and dimensionless times $\tau=t/t_{\rm sf}$. We treat these as free parameters to be fit. In addition, we fit free parameters for each of the models: $t_{\rm acc}$ for model CB, $t_{\rm acc}$ and $\phi_{\rm d}$ for model CBD, $t_{\rm coll}$ for model GC, $t_{\rm coll}$, $t_{\rm fb}$, and $\phi_{\rm d}$ for model GCD, and δ for model IE. Note that we do not have to fit to η or ξ (for the GC and GCD models), because η is absorbed into the definition of $t_{\rm sf}$, and ξ into the definition of $t_{\rm coll}$. We adopt priors that are flat in the logarithm of all the positive-definite quantities (all time-scales) or that are strictly greater than unity ($\phi_{\rm d}$), and flat linear priors in all other parameters. We impose almost no constraint on the time of observation $t_{\rm clust}$, allowing any value in the range 0.01−100 Myr, but we limit the allowed ranges of the remaining parameters based on physical considerations, which we now proceed to describe.

First, for all models we set the prior probability to zero for $\epsilon_{\rm ff}$ outside the range 10^{-4} to 1, on the grounds that $\epsilon_{\rm ff}$ values outside this range correspond to unphysically-inefficient or efficient star formation; to estimate $\epsilon_{\rm ff}$ from $t_{\rm sf}$, we use the observed free-fall time in NGC 6530 or the ONC, as appropriate, and $\eta=1$.[2] This serves to define the allowed range of $t_{\rm sf}$. Second, we apply priors based on the physical picture that motivates each model. For the CB and CBD models, the physical picture is that accretion is due to the collapse of a larger-scale, lower-density reservoir with a longer dynamical time than the cluster-forming region, a picture that requires $t_{\rm acc}>t_{\rm ff}$; we also require $t_{\rm acc}\le t_{\rm clust}$, not for any physical reason, but simply because all models with $t_{\rm acc}>t_{\rm clust}$ have identical age distributions for the stars that exist today, and thus cannot be distinguished in our analysis. For the GC and GCD models, the central idea is that regions collapse on a free-fall time-scale, forming stars while doing so. We therefore impose as a prior $0.1<\xi<10$; lower values of ξ correspond to collapses so slow as to be nearly indistinguishable from the ST model, while higher values require regions to collapse much faster than a free-fall time, which is unphysical. This serves to limit the range of $t_{\rm coll}$ (see footnote 2). Finally, for IE, theoretical models of how the density structure changes as star formation proceeds predict $\delta\approx 1$. We allow some range around this, by setting our prior to zero outside the range $\delta=0-3$.

Our third and final prior is on the present-day SFE, $\epsilon_{*,\rm clust}\equiv M_*(t_{\rm clust})/M_{\rm g,0}$, i.e. the fraction of all the gas available that has been converted to stars; note that $\epsilon_{*,\rm clust}$ may be smaller than the final

[2]Applying this prior to the GC and GCD cases requires some care, because a particular combination of $t_{\rm clust}$, $t_{\rm sf}$, and $t_{\rm coll}$, the parameters to which we are fitting, does not by itself determine a unique value of $\epsilon_{\rm ff}$; instead, one can change $\epsilon_{\rm ff}$ arbitrarily while leaving all these time-scales unchanged by simultaneously changing ξ and $t_{\rm ff,0}$. To determine $\epsilon_{\rm ff}$, we must therefore choose a value of ξ. We can do so by considering two possible scenarios. One is that the cluster in question has not yet reached collapse ($t_{\rm clust}<t_{\rm coll}$), in which case we can fix ξ by demanding that the free-fall time in the model match the observed present-day free-fall time $t_{\rm ff,clust}$ (0.6 Myr for the ONC, 0.5 Myr for NGC 6530, respectively). Re-arranging equation (21), we find that the value of ξ that satisfies this condition is $\xi=2t_{\rm ff,\,clust}/(t_{\rm coll}-t_{\rm clust})$. This in turn breaks the degeneracy and allows us to determine a unique value of $\epsilon_{\rm ff}$. The other possibility is that the cluster as we see it today is after the collapse to singularity ($t_{\rm clust}>t_{\rm coll}$), in which case the free-fall time we measure is a result of the stars rebounding to their current positions post-collapse, and has nothing to do with the free-fall time prior to collapse. In this case ξ is unconstrained by the fit, and we must therefore adopt a value of ξ. For this case, we choose a fiducial value $\xi=1$. Our calculation of the best-fitting model is able to consider both scenarios, since we do not impose any prior on whether $t_{\rm clust}<t_{\rm coll}$ or $t_{\rm clust}>t_{\rm coll}$.

SFE ϵ_* that would be reached as $t_{\rm clust} \rightarrow \infty$. For the ONC, Kim et al. (2019) find that the cluster is virialised and not expanding, which suggests that its SFE could not be too low. We have no direct dynamical evidence that the same is true for NGC 6530, but given its overall similarity with the ONC, this seems likely to be the case for it as well. The requirement that the SFE not be 'too low' is somewhat difficult to quantity: when gas is removed from a protocluster rapidly compared to its dynamical time, loss of more than ≈ 70 per cent of the mass always leads to complete unbinding (Kroupa, Aarseth & Hurley 2001). However, the age distributions in the ONC and NGC 6530 imply that star formation, and presumably mass removal, have been ongoing for significantly longer than a free-fall time, and for sufficiently adiabatic gas removal, stars can remain bound down to arbitrarily small star formation efficiencies. Moreover, in order to match the observation that most stars do not form as part of bound clusters, we require that only a small fraction of the stars remain bound, and thus we do not want the efficiency to be too high. Given our uncertainties, we adopt a relatively mild prior, which disfavours efficiencies below 5 per cent. Formally, we apply a prior $p_{\rm prior}(\epsilon_{*,\rm obs}) \propto \exp\{-[0.05/\min(\epsilon_{*,\rm obs}, 0.05)]^2\}$. For the purpose of calculating ϵ_*, we adopt $\eta = 1$, corresponding to 50 per cent instantaneous SFE, for all models, and a 50 per cent final SFE for all but the CBD and GCD models. By allowing $\epsilon_{*,\,\rm clust}$ to be small compared to ϵ_*, we are allowing for the possibility that the clusters are observed early in the formation process, when $M_g \gg M_*$.

Finally, we note that the ONC is also observed to have a small gas fraction at the present day (Da Rio et al. 2014), $f_{\rm g,clust} \equiv M_{\rm g}(t_{\rm clust})/[(M_{\rm g}(t_{\rm clust}) + M_*(t_{\rm clust})] \ll 1$. While in principle this could serve as an additional prior, we lack quantitative constraints on the gas fraction in NGC 6530, and, with the exception of CBD and GCD, none of our models contains an explicit treatment of gas clearing. For this reason, we will report $f_{\rm g,clust}$ for our fits, but we will not impose any restrictions on it as a prior.

3.1.3 Results

Having defined the likelihood function and priors, we use the package EMCEE (Foreman-Mackey et al. 2013) to perform a Markov Chain Monte Carlo (MCMC) calculation to determine the posterior probability distribution for all the free parameters in each model as compared to the data; for the CB model we consider both a case with our fiducial value, $p = 3$, and one with $p = 0$, as predicted for late stages of collapse by Lee & Hennebelle (2016a, b). For this calculation we use 100 walkers and perform 1000 MCMC steps; visual inspection of the chains indicates that this is more than adequate for convergence. We report the marginalised posterior PDFs, which we derive from the final 800 steps (i.e. we use 200 steps as a burn in period), in Table 3, and show the fits in Fig. 2. We provide full posterior PDF distributions of all variables as Supplementary material (online). In Table 3 we also report three additional. derived quantities for each model, which are helpful in interpreting the results: the SFR per free-fall time $\epsilon_{\rm ff}$, the present-day gas fraction $f_{\rm g,clust}$, and the present-day SFE $\epsilon_{*,\rm clust}$.

Our analysis allows a few immediate conclusions. First, examining Fig. 2, it is clear that the ST and CB ($p = 0$) models provide a poor description of the data in both NGC 6530 and the ONC. The underlying reason is that ST always produces an SFR that is highest at the start of star formation and then tapers; CB with $p = 0$ has an SFR that accelerates with time only weakly. Both models therefore predict a stellar age distribution that is peaked towards the oldest ages, contrary to what we observe. The MCMC attempts to compensate for this effect by favouring large star formation timescales $t_{\rm sf}$, so that as little gas is converted to stars as possible and the SFR falls off due to gas depletion as little as possible; this is also why both models have very high present-day gas fraction $f_{\rm g,clust}$ and very low present-day SFE $\epsilon_{*,\rm clust}$.

The IE model provides a better fit to the data, but in order to do so the fit is driven to values of δ, the acceleration parameter, far from the theoretically-preferred value $\delta = 1$. Indeed, the only reason δ does not go even higher is that our priors do not allow $\delta > 3$. Physically, this is because the model has difficulty producing a star formation history that extends for many free-fall times but also accelerates strongly at late times, unless δ is very large. The existence of a reasonably population of stars with ages approaching $\sim 10 t_{\rm ff}$ requires that the $t_{\rm clust}$ not be too small, but then if δ is close to unity, too much gas is consumed at early times to allow the SFR to accelerate at later times. Thus in order to fit the data, the model requires a much larger value of δ, which more strongly suppresses star formation at early times.

The most successful models are CB, CBD, GC, and GCD. Though none of the models are able to reproduce the full age distribution in great detail, all four produce accelerating star formation that is in reasonable agreement with the observed age distribution, with an accretion time (for CB or CBD) or a collapse time (for GC and GCD) that is nearly equal to the age of the oldest stars present, and to our best estimate for the age of the system as a whole. Given that our model for the uncertainties in stellar age estimates is almost certainly too simplistic, this is probably the best level of agreement for which it is reasonable to hope. The posterior distribution of dimensionless SFE $\epsilon_{\rm ff}$ in these models is extremely broad, mainly because the star formation history is relative insensitive to gas consumption, and instead reflects the accumulation of additional mass (at a rate in good agreement with that predicted by Goldbaum et al. 2011) in CB or CBD, or to the overall increase in the density and thus decrease in the free-fall time in GC or GCD. Interestingly, in the ONC all three models either admit or require that the present-day gas fraction $f_{\rm g,clust}$ be small, consistent with the observations of Da Rio et al. (2014), though we did not explicitly impose this as a prior.

3.2 $\epsilon_{\rm ff}$ from YSO counts

3.2.1 Data set

The next observational test to which we subject our models is the relationship between gas and YSOs in the gaseous objects that are the likely progenitors of star clusters. As discussed in the Section 1, estimates of $\epsilon_{\rm ff}$ based on YSO counts cluster around ≈ 0.01 in all observed star-forming regions, with small scatter. While this would seem to straightforwardly and directly constrain $\epsilon_{\rm ff}$, a number of authors have suggested that this is not the case due to biases introduced by the methodologies of the measurement. For example, Lee et al. (2016) argue that some measurements preferentially select clouds early in their evolution, when, according to Lee et al.'s favoured IE model, $\epsilon_{\rm ff}$ is smaller than its time-averaged value. Similarly, Vázquez-Semadeni et al. (2019) favour a GCD model and argue that estimates of $\epsilon_{\rm ff}$ may be erroneous in clouds because a count of the number of YSOs present implicitly integrates the SFR over some period of time into the past, when the free-fall time was longer than the value we measure at the present day. We are in a position to test both these hypotheses, by directly modelling the observed distribution of $\epsilon_{\rm ff}$ values produced by our cluster formation models.

Table 3. Best-fitting parameters obtained by comparing each model to the observed distribution of stellar ages in the ONC and NGC 6530. The quantities listed as 'Fit parameters' are those directly constrained in the fit, while 'Derived parameters' are calculated from the fit parameters. Values are specified in the form $p(50)^{p(84)-p(50)}_{p(16)-p(50)}$, where $p(q)$ is the qth percentile of the marginalised posterior PDF for that quantity. Times expressed as logarithms are in Myr.

Model	Fit parameters			Derived parameters		
	$\log t_{\rm clust}$ (Myr)	$\log t_{\rm sf}$ (Myr)	Other	$\log \epsilon_{\rm ff}$	$f_{\rm g,clust}$	$\log \epsilon_{*,\rm clust}$
			ONC			
ST	$0.78^{+0.02}_{-0.02}$	$1.72^{+0.16}_{-0.14}$	–	$-2.24^{+0.14}_{-0.16}$	$0.94^{+0.02}_{-0.02}$	$-1.26^{+0.13}_{-0.15}$
CB, $p=0$	$0.92^{+0.03}_{-0.02}$	$1.46^{+0.17}_{-0.24}$	$\log t_{\rm acc} = 0.89^{+0.03}_{-0.04}$	$-1.98^{+0.24}_{-0.17}$	$0.92^{+0.02}_{-0.05}$	$-1.15^{+0.21}_{-0.15}$
CB, $p=3$	$1.13^{+0.04}_{-0.04}$	$-0.28^{+1.33}_{-0.19}$	$\log t_{\rm acc} = 1.12^{+0.04}_{-0.04}$	$-0.24^{+0.19}_{-1.33}$	$0.12^{+0.73}_{-0.07}$	$-0.33^{+0.02}_{-0.55}$
CBD, $p=3$	$1.12^{+0.04}_{-0.07}$	$0.11^{+1.35}_{-0.47}$	$\log t_{\rm acc} = 1.11^{+0.04}_{-0.17}$, $\log \phi_{\rm d} = 1.48^{+0.38}_{-1.10}$	$-0.63^{+0.47}_{-1.35}$	$0.02^{+0.89}_{-0.02}$	$-0.44^{+0.09}_{-0.67}$
GC	$0.94^{+0.04}_{-0.04}$	$1.34^{+0.78}_{-0.11}$	$\log t_{\rm coll} = 0.94^{+0.07}_{-0.05}$, $\xi = 1.00^{+0.18}_{-0.06}$	$-1.00^{+0.07}_{-0.72}$	$0.00^{+0.92}_{+0.00}$	$-0.30^{+0.00}_{-0.82}$
GCD	$0.99^{+0.05}_{-0.05}$	$1.92^{+0.43}_{-0.49}$	$\log t_{\rm coll} = 0.99^{+0.05}_{-0.05}$, $\log t_{\rm fb} = 0.81^{+0.15}_{-0.64}$, $\log \phi_{\rm d} = 0.60^{+0.53}_{-0.44}$, $\xi = 1.00^{+2.89}_{+0.00}$	$-1.33^{+0.31}_{-0.46}$	$0.00^{+0.64}_{+0.00}$	$-0.74^{+0.35}_{-0.42}$
IE	$1.08^{+0.04}_{-0.04}$	$4.76^{+0.47}_{-0.61}$	$\delta = 2.68^{+0.23}_{-0.42}$	$-1.31^{+0.09}_{-0.08}$	$0.92^{+0.03}_{-0.06}$	$-1.13^{+0.23}_{-0.16}$
			NGC 6530			
ST	$0.52^{+0.02}_{-0.02}$	$1.64^{+0.13}_{-0.16}$	–	$-2.25^{+0.16}_{-0.13}$	$0.96^{+0.01}_{-0.02}$	$-1.44^{+0.15}_{-0.12}$
CB, $p=0$	$0.66^{+0.02}_{-0.02}$	$1.35^{+0.18}_{-0.12}$	$\log t_{\rm acc} = 0.64^{+0.03}_{-0.03}$	$-1.95^{+0.12}_{-0.18}$	$0.95^{+0.02}_{-0.02}$	$-1.29^{+0.11}_{-0.16}$
CB, $p=3$	$0.92^{+0.02}_{-0.03}$	$1.03^{+0.23}_{-0.39}$	$\log t_{\rm acc} = 0.90^{+0.03}_{-0.03}$	$-1.63^{+0.39}_{-0.23}$	$0.91^{+0.03}_{-0.11}$	$-1.10^{+0.32}_{-0.18}$
CBD, $p=3$	$0.88^{+0.03}_{-0.27}$	$0.52^{+1.04}_{-0.36}$	$\log t_{\rm acc} = 0.87^{+0.03}_{-0.97}$, $\log \phi_{\rm d} = 1.84^{+0.12}_{-1.59}$	$-1.13^{+0.36}_{-1.04}$	$0.43^{+0.53}_{-0.34}$	$-0.77^{+0.20}_{-0.61}$
GC	$0.83^{+0.04}_{-0.05}$	$2.23^{+0.20}_{-1.03}$	$\log t_{\rm coll} = 0.86^{+0.04}_{-0.09}$, $\xi = 2.64^{+0.67}_{-1.64}$	$-1.50^{+0.47}_{-0.21}$	$0.93^{+0.02}_{-0.93}$	$-1.21^{+0.91}_{-0.17}$
GCD	$0.87^{+0.04}_{-0.04}$	$2.40^{+0.18}_{-0.17}$	$\log t_{\rm coll} = 0.88^{+0.04}_{-0.04}$, $\log t_{\rm fb} = 0.58^{+0.19}_{-0.67}$, $\log \phi_{\rm d} = 0.87^{+0.28}_{-0.22}$, $\xi = 8.20^{+1.34}_{-7.20}$	$-1.28^{+0.19}_{-0.82}$	$0.89^{+0.04}_{-0.89}$	$-1.29^{+0.14}_{-0.17}$
IE	$0.85^{+0.02}_{-0.02}$	$4.60^{+0.20}_{-0.20}$	$\delta = 2.95^{+0.04}_{-0.08}$	$-1.14^{+0.04}_{-0.05}$	$0.95^{+0.02}_{-0.02}$	$-1.29^{+0.11}_{-0.17}$

Note. We derive $\epsilon_{\rm ff}$ as follows: for models ST, CB, and CBD, we use equation (4) with $t_{\rm ff}$ set equal to the observed value in NGC 6530 or the ONC. For model IE, we report the time-averaged value $\bar{\epsilon}_{\rm ff}$ given by equation (33). Finally, for models GC and GCD we use the procedure described in footnote 2. In all cases our numerical value is for $\eta = 1$, and $\epsilon_{\rm ff}$ obeys the scaling $\epsilon_{\rm ff} \propto 1/(1+\eta)$.

We take our measured distribution of $\epsilon_{\rm ff}$ values from Heyer et al. (2016), who identify class 0/I YSOs within and measure $\epsilon_{\rm ff}$ for gas clumps identified in the ATLASGAL survey (Schuller et al. 2009; Csengeri et al. 2014). We use Heyer et al.'s IMF-corrected estimates of $\epsilon_{\rm ff}$, which account statistically for the fact that their YSO catalogues begin to suffer from incompleteness for protostars smaller than ≈ 2 M$_\odot$. This is the largest ($N = 517$)[3] and most complete sample of $\epsilon_{\rm ff}$ measurements in the literature, and the ATLASGAL clumps that it targets are very similar to the ONC and NGC 6530 in terms of mass, density, and free-fall time, making the data well-suited to the task of using both the cluster star formation history and the $\epsilon_{\rm ff}$ distribution together, as we do below. Specifically, the mean free-fall time of the ATLASGAL clumps is 0.3 Myr, very similar to the observed free-fall times of 0.5 and 0.6 Myr in NGC 6530 and the ONC. Thus the ATLASGAL sample very likely represents a survey of YSOs in objects that are will become clusters like NGC 6530 or the ONC, just at a slightly earlier evolutionary phase. However, we do note that the distribution of $\epsilon_{\rm ff}$ values obtained by Heyer et al. (2016) is qualitatively quite similar to those obtained from other samples that also use YSO counts for objects at a range of size and density scales (e.g. Evans et al. 2014; Ochsendorf et al. 2017).

[3]For some of this sample, Heyer et al. (2016) do not detect any YSOs, and thus only obtain an upper limit on $\epsilon_{\rm ff}$. For the purposes of our analysis, we take the value of $\epsilon_{\rm ff}$ in these clumps to be equal to the stated 2σ upper limit.

3.2.2 *Likelihood function*

As in Section 3.1, to compare to the models to the observations we require a likelihood function that gives the probability density of the data given the model, which must properly account for averaging of the SFR over a finite time interval, potential biases in the sample, and observational errors. First consider the issue of averaging over a finite time. The Heyer et al. (2016) data set on which we focus estimates the SFR based on number counts of class 0/I YSOs, a phase that lasts for a time $t_{\rm YSO} \approx 0.5$ Myr (Evans et al. 2009; Gutermuth et al. 2009). We can therefore define an appropriately time-averaged $\epsilon_{\rm ff}$ for our models as

$$\epsilon_{\rm ff,avg}(t, \Delta t) = \frac{\left[M_*(t) - M_*(t-\Delta t)\right]/\Delta t}{M_{\rm g}(t)/t_{\rm ff}(t)}, \tag{39}$$

where t is the time of observation, and $\Delta t = 0.5$ Myr is the window over which the SFR is averaged.

As with our treatment of stellar ages, we must consider not only biases (in this case introduced by averaging over a finite time), but observational errors. Errors in $\epsilon_{\rm ff}$ measurements are significantly more poorly modelled than errors in stellar age distributions, and involve subtleties such as making an IMF-based correction for the presence of protostars too dim to be detected. Given our ignorance, we will adopt a simple lognormal functional form, i.e. in a cloud with a true (time-averaged) logarithmic star formation efficiency $\log \epsilon_{\rm ff,avg}$, the distribution of observationally inferred values $\log \epsilon_{\rm ff,obs}$ will be distributed as a Gaussian of width $\sigma_{\log \epsilon_{\rm ff}}$ centred on $\log \epsilon_{\rm ff,avg}$. That is, given a true (time-averaged) efficiency

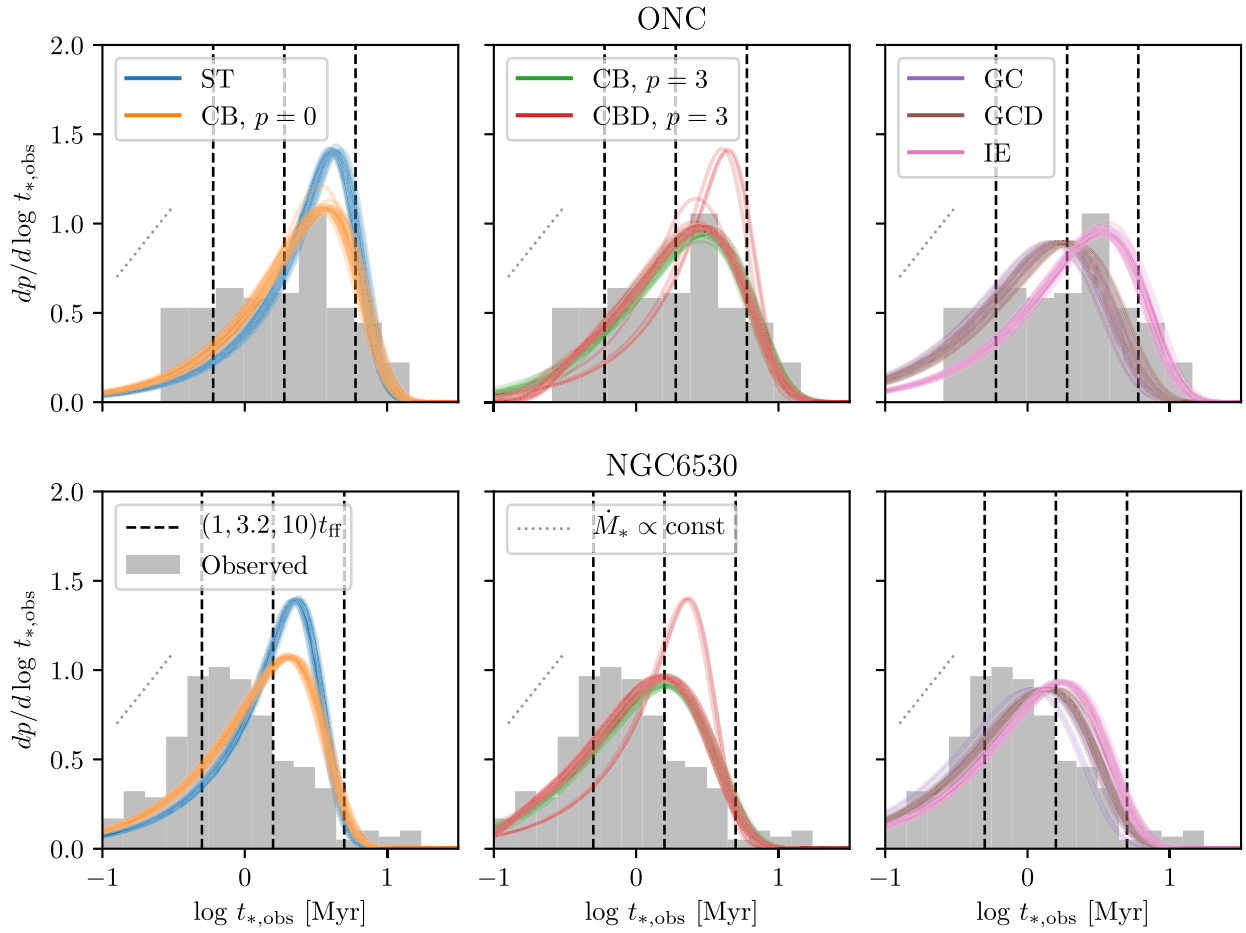

Figure 2. Distribution of observed stellar ages dp/dlog $t_{*,\rm obs}$ in the ONC (top) and NGC 6530 (bottom). In all panels, the coloured lines represent 20 random samples from the final iteration of the MCMC for each of the models, as indicated in the legend. Grey histograms show the observed distribution, and are the same in every panel. The dashed vertical lines indicate 1, 3.2, and 10× the observed free-fall time, as indicated. The dotted grey lines, which we provide to guide the eye, are lines of slope unity, which, given the logarithmic age bins, corresponds to a constant SFR.

per free-fall time $\epsilon_{\rm ff,avg}$, the distribution of observationally estimated star formation efficiency per free-fall time is

$$f(\log\epsilon_{\rm ff,obs} \mid \log\epsilon_{\rm ff,avg}) = \frac{1}{\sqrt{2\pi}\sigma_{\log\epsilon_{\rm ff}}} \exp\left[-\frac{\left(\log\epsilon_{\rm ff,avg} - \log\epsilon_{\rm ff,obs}\right)^2}{2\sigma^2_{\log\epsilon_{\rm ff}}}\right]. \quad (40)$$

The value of the dispersion $\sigma_{\log\epsilon_{\rm ff}}$ is not well known, but we will see below that it is not necessary to adopt a model for $\sigma_{\log\epsilon_{\rm ff}}$; instead we can leave $\sigma_{\log\epsilon_{\rm ff}}$ as a parameter to be fit along with other model parameters.

Now consider a cloud observed at some time t during its evolution, with an instantaneous time-averaged star formation efficiency $\epsilon_{\rm ff,avg}(t, \Delta t)$. The distribution of observed efficiencies for this cloud is $f(\log\epsilon_{\rm ff,obs}|\log\epsilon_{\rm ff,avg}(t, \Delta t))$. If we have a population of such clouds, each observed at random times t between the onset of star formation at $t = 0$ and some maximum time $t_{\rm max}$, then the distribution of observed $\epsilon_{\rm ff,obs}$ values for the population is simply the average of $f(\log\epsilon_{\rm ff,obs}|\log\epsilon_{\rm ff,avg}(t, \Delta t))$ over all possible times t at which the clouds could be observed, i.e.

$$\frac{{\rm d}p}{{\rm d}\log\epsilon_{\rm ff,obs}} = \frac{1}{t_{\rm max}}\int_0^{t_{\rm max}} f(\log\epsilon_{\rm ff,obs} \mid \log\epsilon_{\rm ff,avg}(t, \Delta t))\,{\rm d}t. \quad (41)$$

The choice of maximum time $t_{\rm max}$ is somewhat subtle. In simple models where $M_{\rm g}$ reaches 0 in finite time, such as the GC model, one can simply take $t_{\rm max}$ to be the time for which $M_{\rm g}(t_{\rm max}) = 0$. However, we are interested in comparing to a more general class of models where $M_{\rm g}$ may not go to exactly 0 at finite time. To choose a reasonable $t_{\rm max}$, we note that studies of $\epsilon_{\rm ff}$ based on YSO counts always select YSOs and gas clouds within the same area on the sky, which limits the phase of evolution to which they are sensitive: as clusters evolve and begin to clear their gas, stars inevitably cease to be surrounded by molecular gas, so clouds that have cleared most of their gas are not included in YSO counting surveys. Our simple zero-dimensional models cannot capture this effect directly, but we crudely mimic it by choosing our time interval to correspond to that over which $M_{\rm g}/M_* > 1$, i.e. when the stellar mass has not yet exceeded the gas mass. We therefore take $t_{\rm max}$ to be defined implicitly by the condition $M_{\rm g}(t_{\rm max})/M_*(t_{\rm max}) = 1$. We have verified that varying the value of $M_{\rm g}/M_*$ we use to define our time interval by a factor of ten in either direction not change the results substantially.

Given the preceding discussion, we have now write down the log likelihood function for a set of observed $\epsilon_{\rm ff}$ values is

$$\log\mathcal{L} = \sum_{i=1}^{N} \log\left(\frac{{\rm d}p}{{\rm d}\log\epsilon_{\rm ff,obs}}\right)_{\epsilon_{\rm ff,obs}=\epsilon_{{\rm ff},i}}, \quad (42)$$

where $\epsilon_{\rm ff,}{}_i$ is the ith observed value of $\epsilon_{\rm ff}$, and there are N measurements in total. We use this likelihood function with EMCEE to obtain posterior PDFs for the parameters for the same models as in Section 3.1. As in our analysis of the stellar age distribution, we use priors that are flat in the logarithm of positive-definite quantities, and flat in value for other quantities; the allowed parameter range is identical to that used in Section 3.1. In addition to the parameters included there, we must also fit for η, $\sigma_{\log\epsilon_{\rm ff}}$, and ξ (for model GC and GCD), since, while these do not affect the distribution of stellar ages, they do affect the distribution of observed $\epsilon_{\rm ff}$ values. For η our prior is flat in log from 0.01 to 10, and for $\sigma_{\log\epsilon_{\rm ff}}$ it is flat in log from 0.01 to 10. We must also choose a value for the free-fall time, since this sets the ratio $\Delta t/t_{\rm ff}$, which determines how much the observed $\epsilon_{\rm ff}$ distribution is biased by averaging the SFR over a finite time. As noted above, the mean value of $t_{\rm ff}$ in the ATLASGAL sample is 0.3 Myr, and the dispersion around this is small (0.26 dex), so we use $t_{\rm ff} = 0.3$ Myr for our analysis of all models except GC and GCD; these models sweep through all values of $t_{\rm ff}$ from $t_{\rm ff,0}$ to 0, so for this case we impose as a prior the requirement that $t_{\rm ff,0} >$ 0.3 Myr, i.e. the collapse must start from a state that is no denser than the observed ATLASGAL clumps.

3.2.3 Results

We show models evaluated using samples drawn from the MCMC chains in Fig. 3, and report the posterior PDFs of all parameters in Table 4. The results show that all the models we consider can fit the observed $\epsilon_{\rm ff}$ distribution quite well, but that both $\epsilon_{\rm ff}$ and the level of observational error are *very* tightly constrained by the observations; $\epsilon_{\rm ff}$ is required to be of order a few per cent, and $\sigma_{\log\epsilon_{\rm ff}}$ to be approximately 0.15 dex. Indeed, the models even constrain η not be too large, since otherwise rapid mass removal means that the gas mass is able to change significantly over the time-averaging interval Δt, which in turn would broaden the observed $\epsilon_{\rm ff}$ distribution more than the data allow. Thus, despite the hypothesis in the literature that measured $\epsilon_{\rm ff}$ distributions are biased because they average over a finite time interval and thus miss changes in the free-fall time (e.g. Vázquez-Semadeni et al. 2019), or that they miss periods of efficient star formation (e.g. Lee et al. 2016), we do not obtain significantly looser constraints on the value of $\epsilon_{\rm ff}$ when we explicitly put those possibilities into our model.

3.3 Combined constraints

Having examined the constraints we can deduce from the distribution of stellar ages and the YSO–gas correlation individually, we now ask whether these constraints are compatible. That is, do there exist a set of parameters for a given model such it can simultaneously reproduce the observed stellar age distribution in young clusters and the YSO count in protoclusters? To answer this question, we use our MCMC samples to compute the dimensionless parameters – $\epsilon_{\rm ff}$, η, etc. – that characterize each proposed model, using the constraints from both the stellar age distribution and YSO counts. We focus only on the dimensionless parameters, since, while the star clusters for which we have examined the stellar age distribution and the ATLASGAL clumps are similar in terms of mass and free-fall time, they are not completely identical, and thus we do not expect the dimensional parameters (e.g. free-fall time or collapse time) to match exactly. We plot the posterior PDFs of the dimensionless parameters for models CBD, GCD, and IE in Figs 4, 5, and 6, respectively.[4] These plots use the posterior PDFs derived from the stellar age distribution in NGC 6530, since it is a somewhat larger data set, but the results for the ONC are qualitatively similar. We omit ST and CB ($p = 0$) from this comparison because we have already determined that these models provide poor fits to the stellar age distribution alone, and we omit CB ($p = 3$) and GC because they are qualitatively similar to CBD and GCD, respectively, on the parameters they share. However, the corresponding plots for these clusters are provided in the Supplementary material (online).

Turning first to Fig. 6, we immediately see that the IE model has a major difficulty: as discussed in Section 3.1 and shown in Fig. 6, the stellar ages distributions in NGC 6530 and the ONC are best fit in the context of this model by a star formation efficiency that increases as roughly $\epsilon_{\rm ff} \propto t^3$ (or faster, since $\delta = 3$ is the largest allowed by our priors). This is completely at odds with the constraint provided by the ATLASGAL clumps, whose tight relationship between YSOs and gas properties requires that $\epsilon_{\rm ff}$ be nearly constant, and thus that $\delta \approx 0$. The physical explanation for this tension is simple: star formation is observed to accelerate based on stellar age distributions, and the IE model interprets this acceleration as a systematic increase in star formation efficiency with time. However, when one observes the gas clumps that are in the process of forming clusters, one finds that the number of YSOs per unit gas mass, normalised by the free-fall time, is nearly constant, completely inconsistent with large variations in star formation efficiency. There is no way to reconcile these two constraints in the context of the IE model, or indeed in any model that assumes the acceleration of star formation is due to an increase in star formation efficiency with time. Instead, the acceleration of star formation must be due either to an increase in the star-forming mass with time (as in CB or CBD) or a decrease in the free-fall time (as in GC or GCD). We may therefore rule out the IE model.

The CBD and CGD models illustrated respectively in Figs 4 and 5, on the other hand, show no contradiction between the parameter values demanded by the stellar age distributions and the ATLASGAL clumps. In both sets of models the ATLASGAL data very tightly constrain $\epsilon_{\rm ff}$, while setting little constraint on any other parameters. Conversely, the stellar age distribution tightly constrains $\tau_{\rm acc}$, $\tau_{\rm coll}$, ξ, $x_{\rm fb}$, and $\phi_{\rm d}$, but provides little restriction on $\epsilon_{\rm ff}$. As a result, there is a reasonable parameter space of overlap.

Thus we find that the joint set of data favour one of two scenarios. We plot the history of gas and stellar mass, SFR, and mean density and free-fall time derived for these two scenarios in Fig. 7. In the first, gas accretes as roughly $\dot{M} \propto t^3$ (consistent with the theoretical models of Goldbaum et al. 2011) and forms stars inefficiently ($\epsilon_{\rm ff} \approx 0.01$). Accretion continues for ∼1–10 star formation time-scales ($\tau_{\rm acc} \sim$ 1–10), and once it ends, mass is rapidly dispersed by feedback ($\phi_{\rm d} \gg 1$). The precise parameters used for the CBD model shown in Fig. 7 are $\log\epsilon_{\rm ff} = -1.75$, $\tau_{\rm acc} = 3$, $\eta = 3$, $\phi_d = 10$; all of these parameters are within the 16th to 84th percentile range allowed by both sets of constraints. The gas and stellar masses in the model, physical time, and SFR, can be rescaled arbitrarily by changing the total cloud mass and density, while leaving all the dimensionless parameters (which

[4]For parameters that cannot be constrained by the stellar age distribution, we take the posterior PDF derived from stellar ages to be equal to the flat prior we use for these variables when analysing the ATLASGAL data.

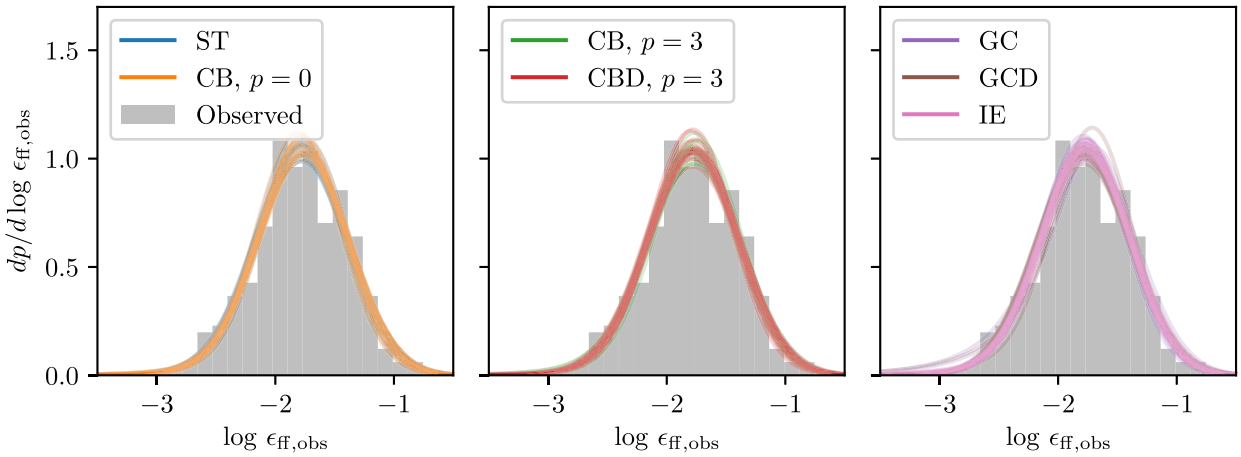

Figure 3. Distribution of observed star formation efficiencies $\log\epsilon_{\rm ff,obs}$. Grey histograms show the distribution observed by Heyer et al. (2016) for the ATLASGAL sample, and are the same in every panel. Coloured lines represent 20 random samples from the final iteration of the MCMC fit for each model, as indicated in the legend.

Table 4. Best-fitting parameters obtained by comparing each model to the observed distribution of measured $\epsilon_{\rm ff}$ values in ATLASGAL clumps (Heyer et al. 2016).

Model	Fit parameters				Derived parameters
	$\log\sigma_{\log\epsilon_{\rm ff}}$ (dex)	$\log\eta^a$	$\log t_{\rm sf}$ (Myr)	Other	$\log\epsilon_{\rm ff}^b$
ST	$-0.79^{+0.01}_{-0.01}$	$-0.72^{+0.86}_{-0.86}$	$1.18^{+0.06}_{-0.30}$	–	$-1.78^{+0.02}_{-0.02}$
CB, $p = 0$	$-0.79^{+0.02}_{-0.02}$	$-0.53^{+0.82}_{-0.93}$	$1.13^{+0.10}_{-0.36}$	$\log t_{\rm acc} = 1.45^{+0.38}_{-0.81}$	$-1.76^{+0.02}_{-0.02}$
CB, $p = 3$	$-0.79^{+0.02}_{-0.02}$	$-0.65^{+1.02}_{-0.89}$	$1.13^{+0.08}_{-0.44}$	$\log t_{\rm acc} = 1.77^{+0.16}_{-0.28}$	$-1.74^{+0.02}_{-0.02}$
CBD, $p = 3$	$-0.79^{+0.02}_{-0.02}$	$-0.66^{+0.92}_{-0.88}$	$1.13^{+0.08}_{-0.38}$	$\log t_{\rm acc} = 1.78^{+0.15}_{-0.26}$, $\log\phi_{\rm d} = 1.02^{+0.68}_{-0.70}$	$-1.74^{+0.03}_{-0.02}$
GC	$-0.78^{+0.02}_{-0.02}$	$-0.08^{+0.70}_{-0.82}$	$2.53^{+0.89}_{-0.86}$	$\log t_{\rm coll} = 1.58^{+0.31}_{-0.50}$, $\log\xi = -0.00^{+0.72}_{-0.66}$	$-1.77^{+0.03}_{-0.02}$
GCD	$-0.79^{+0.02}_{-0.03}$	$0.17^{+0.54}_{-0.65}$	$2.47^{+0.72}_{-0.88}$	$\log t_{\rm coll} = 1.39^{+0.43}_{-0.57}$, $\log\xi = 0.22^{+0.52}_{-0.72}$	$-1.77^{+0.06}_{-0.03}$
IE	$-0.79^{+0.02}_{-0.02}$	$-0.69^{+0.79}_{-0.88}$	$1.25^{+0.12}_{-0.27}$	$\log t_{\rm fb} = -0.45^{+1.02}_{-1.04}$, $\log\phi_{\rm d} = 0.82^{+0.67}_{-0.57}$, $\delta = 0.06^{+0.09}_{-0.04}$	$-1.76^{+0.03}_{-0.03}$

Notes. Formatting is identical to that used in Table 3.
[a] The median and percentile values we report for $\log\eta$ are strongly affected by our prior $\log\eta > -2$. All models with $\eta \ll 1$ are essentially identical, so our analysis cannot distinguish them; thus the values we report should be read as providing an upper limit at the reported 84th percentile, rather than a meaningful central estimate.
[b] The value of $\epsilon_{\rm ff}$ we report here is the true value defined by the instantaneous SFR, not the time-averaged value $\epsilon_{\rm ff,avg}$ defined by equation (39). For model IE, we report the time-averaged value $\bar{\epsilon}_{\rm ff}$ given by equation (33). We compute $\epsilon_{\rm ff}$ as described in the notes to Table 3.

determine the shape of the curves) fixed. We have scaled the curves shown to values typical of NGC 6530 and the ONC, and of the ATLASGAL clumps: a final stellar mass of 2000 $M_\odot$, and a free-fall time of 0.3 Myr. The corresponding physical star formation and accretion time-scales are $t_{\rm sf} = 4.2$ Myr and $t_{\rm acc} = 12.7$ Myr, respectively.

In the second scenario, an initially low-density cloud undergoes a global collapse that is fairly rapid compared to the instantaneous free-fall time ($\xi \gtrsim 1$), as might be expected for example in a colliding flow where the collapse is due to external pressure plus gravity rather than gravity alone, but during this collapse it forms stars quite inefficiently ($\epsilon_{\rm ff} \approx 0.01$). As a result, the total collapse time is quite small compared to the star formation time-scale ($\tau_{\rm coll} \lesssim 0.1$), so that most stars form only during the final plunge when the density is running way to infinity – a value $\tau_{\rm coll} < 1$ is required to yield an accelerating star formation history. The plot shown in Fig. 7 uses $\log\epsilon_{\rm ff} = -1.75$, $\tau_{\rm coll} = 0.04$, $\tau_{\rm fb} = 0.036$, $\eta = 1.0$ (so $\xi = 1.8$), and $\phi_{\rm d} = 10$, together with an initial free-fall time $t_{\rm ff,0} = 10$ Myr, again falling within the 16th–84th percentile range of our analysis of NGC 6530 and the ONC; the mass has also been scaled to produce a final stellar mass of 2000 $M_\odot$. The corresponding initial star formation and collapse time-scales are $t_{\rm sf} = 281$ Myr and $t_{\rm coll} = 11.2$ Myr, respectively; the collapse time-scale corresponds to a starting density $\approx$20 cm^{-3}, and thus typical of the cold neutral medium (CNM). In this model, the ATLASGAL clouds began their lives as clouds of CNM, and their present-day properties would correspond to a physical state near the point where the blue and orange lines cross in the bottom panel of Fig. 7.

3.4 Global SFR

We now add an additional constraint to our modelling: the SFR of the Milky Way as a whole is $\approx$2 $M_\odot$ yr^{-1} (Chomiuk & Povich 2011), so the total SFR implied by a successful model must not exceed this value. To see what this implies, we again return to the

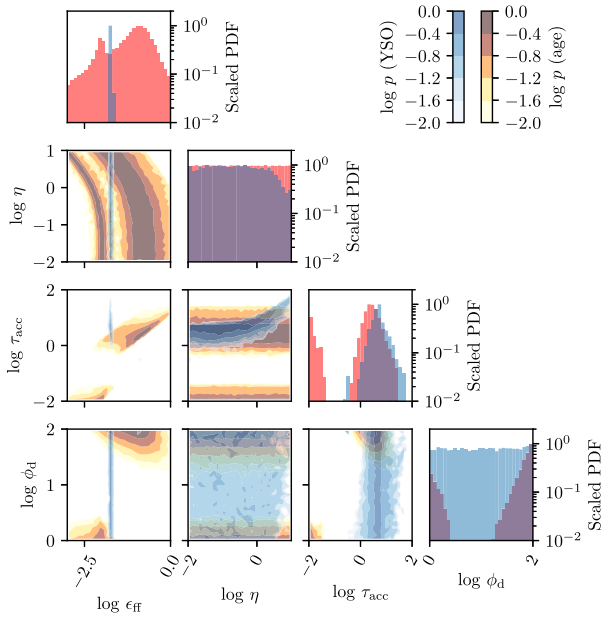

Figure 4. Corner plot showing the posterior PDF for the dimensionless parameters of the CBD model ($\epsilon_{\rm ff}$, η, $\tau_{\rm acc}$, and $\phi_{\rm d}$), derived using the distribution of stellar ages in NGC 6530 (red colours) and the counts of YSOs in ATALASGAL clumps (blue colours). In the panels on the bottom left corner, contours show 2D marginal posterior PDFs for each combination of variables, as indicated on the axes. Histograms along the central diagonal show 1D marginal posterior PDFs for each variable. PDFs in all panels are scaled so that the maximum is unity.

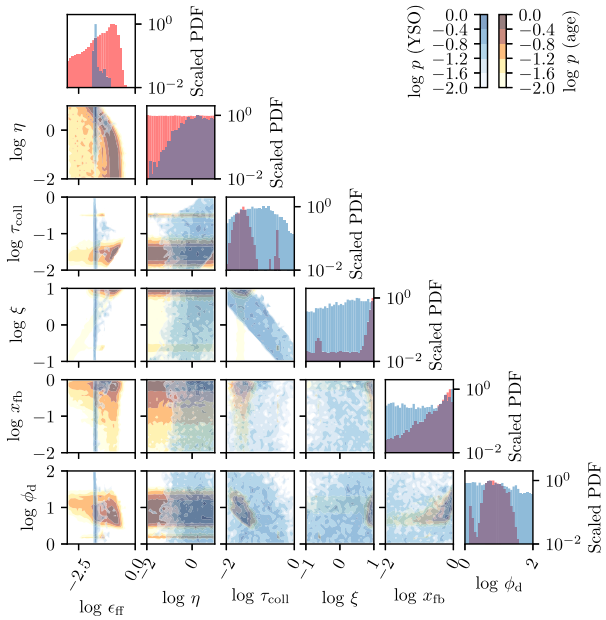

Figure 5. Same as Fig. 4, but showing the GCD model and its dimensionless parameters. Note that we show $x_{\rm fb} = t_{\rm fb}/t_{\rm coll}$ rather than $\tau_{\rm fb} = t_{\rm fb}/t_{\rm sf}$, because the former quantity is more helpful for the discussion that follows.

ATLASGAL sample. As noted above, the mean free-fall time of these objects is $t_{\rm ff} = 0.3$ Myr (Heyer et al. 2016), and the total mass of ATLASGAL clumps in the Galaxy is $M_{\rm tot} \approx 1.0 \times 10^7$ M$_\odot$ (Urquhart et al. 2018).

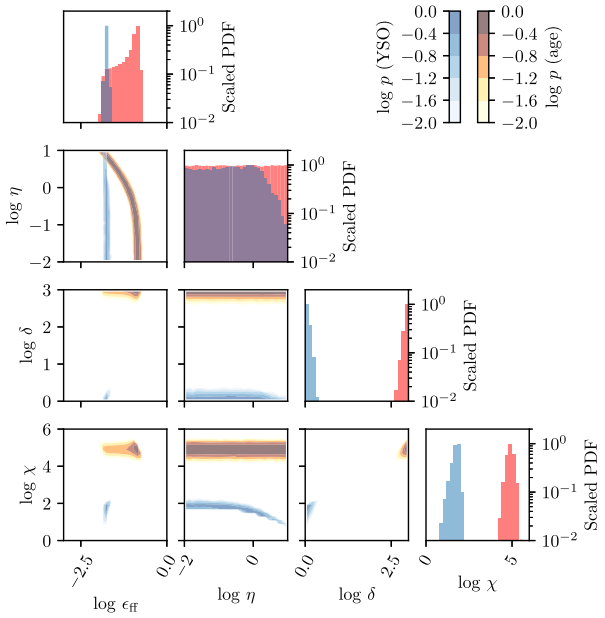

Figure 6. Same as Fig. 4, but showing the IE model and its dimensionless parameters.

3.4.1 ST, CB, and CBD

The rate at which ATLASGAL clumps form stars is straightforward to calculate in the ST, CB, and CBD models:

$$\mathrm{SFR} = \epsilon_{\rm ff}\frac{M_{\rm tot}}{t_{\rm ff}} = 0.33\left(\frac{\epsilon_{\rm ff}}{0.01}\right)\left(\frac{M_{\rm tot}}{10^7\,\mathrm{M}_\odot}\right)\left(\frac{t_{\rm ff}}{0.3\,\mathrm{Myr}}\right)^{-1}\mathrm{M}_\odot\,\mathrm{yr}^{-1}, \quad (43)$$

where we have normalized to the mean free-fall time for the ATLASGAL clumps. Thus if $\epsilon_{\rm ff} \approx 0.01$ for these models, as suggested by our analysis so far, the total contribution of the ATLASGAL clumps to the total star formation budget of the Milky Way is $\approx$0.3 M$_\odot$ yr^{-1}, which is $\approx$ 10 per cent of the total. This is consistent with the upper limit stated above, and in fact suggests a nice consistency: the ATLASGAL clumps are much denser than the mean star-forming region or star cluster (for example, compare to fig. 9 of Krumholz et al. 2019), and thus the stars that form within them are much more likely to remain part of a bound cluster than the typical star formed in the Galaxy. If we hypothesize that the ATLASGAL clumps correspond roughly to the bound portion of the star formation in the Galaxy, so our estimate implies that $\sim$ 10 per cent of all stars formed in bound clusters, that is entirely consistent with the observationally measured fraction of stars formed in bound clusters in typical spiral galaxies (e.g. Ryon et al. 2014; Adamo et al. 2015; Johnson et al. 2016; Chandar et al. 2017). We caution, however, not to put too much weight on this agreement, since we do not in fact know if the density range that is selected by ATLASGAL corresponds well to the conditions that delineate between bound and unbound star formation.

3.4.2 GC, GCD, and IE

The remaining models require a more refined treatment because $t_{\rm ff}$ and $\epsilon_{\rm ff}$ can vary. Since these models do not depend on the magnitude of the mass, we can assume that the entire population of ATLASGAL clouds is born with the same mass and then evolves

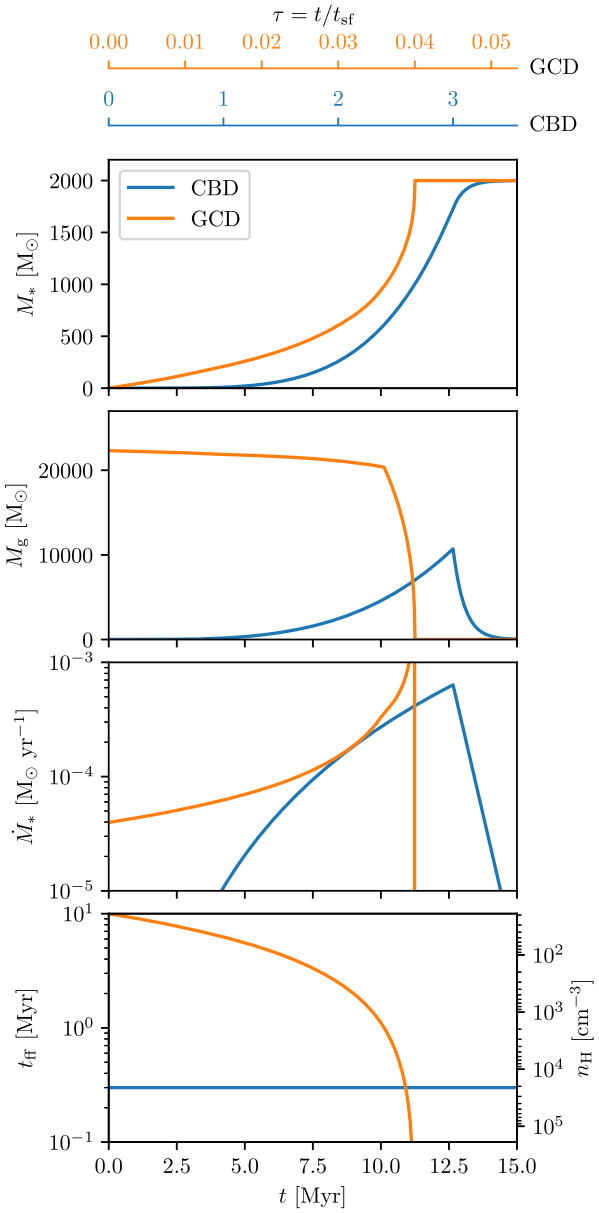

Figure 7. Example histories of stellar mass, gas mass, SFR, and free-fall time / density for the two best-fitting models, CBD and GCD, scaled to mass and time-scales typical of the ATLASGAL sample; the exact parameters used to construct these models are described in Section 3.3. The right axis in the bottom panel shows number density of H nuclei, computed assuming a mean mass of $1.4m_{\rm H}$ per H nucleon. The bottom horizontal axis shows physical time in Myr, while the top two axes show dimensionless time $\tau = t/t_{\rm sf}$; this is different for the CBD and GCD models because the star formation time-scale $t_{\rm sf}$ is different in the two models.

according to one of these models. Let $\mathcal{N}_M = {\rm d}\mathcal{N}/{\rm d}M_{\rm g}$ be the number of clouds per unit mass, and let $\dot{\mathcal{N}}$ be the rate at which clouds are born with a mass $M_{\rm g,0}$. The equation of continuity for the cloud mass distribution is then

$$\frac{\partial \mathcal{N}_M}{\partial t} + \frac{\partial}{\partial M_{\rm g}}\left(\mathcal{N}_M \dot{M}_{\rm g}\right) = \dot{\mathcal{N}}\delta(M_{\rm g} - M_{\rm g,0}), \tag{44}$$

so that in a steady state we have

$$-\mathcal{N}_M \dot{M}_{\rm g} = \dot{\mathcal{N}}. \tag{45}$$

The SFR is then

$${\rm SFR} = \int_0^{M_{\rm g,0}} \dot{M}_* \mathcal{N}_M dM = \dot{\mathcal{N}}\epsilon_* M_{\rm g,0} \tag{46}$$

from equation (2), since there is no accretion in these models ($\dot{M}_{\rm acc} = 0$). Here ϵ_* is the final star formation efficiency: $1/(1+\eta)$ in GC or IE, and the value given by equation (30) for GCD. Now the total mass in clouds is

$$M_{\rm tot} = \int_0^{M_{\rm g,0}} M_{\rm g}\mathcal{N}_M \, {\rm d}M, \tag{47}$$

$$= \int_0^{\infty} M_{\rm g}\left(\frac{\dot{\mathcal{N}}}{\dot{M}_{\rm g}}\right)\left(\frac{{\rm d}M_{\rm g}}{{\rm d}t}\right){\rm d}t, \tag{48}$$

$$= \dot{\mathcal{N}} t_{\rm sf}\int_0^{\infty} M_{\rm g}\,{\rm d}\tau. \tag{49}$$

The SFR per unit gas mass is then

$$\frac{\rm SFR}{M_{\rm tot}} = \frac{\epsilon_*}{t_{\rm sf}}\left[\int_0^{\infty}\frac{M_{\rm g}}{M_{\rm g,0}}\,{\rm d}\tau\right]^{-1}. \tag{50}$$

We can check this by noting that for the ST model it gives the result in equation (1):

$${\rm SFR} = \frac{M_{\rm tot}}{(1+\eta)t_{\rm sf}} = \epsilon_{\rm ff}\frac{M_{\rm tot}}{t_{\rm ff}}. \tag{51}$$

First consider the IE model. Evaluating the integral in equation (50) with the aid of equation (32) we find

$${\rm SFR} = \frac{\chi^{\delta/(1+\delta)}}{(1+\delta)^{1/(1+\delta)}\Gamma\left(1+\frac{1}{1+\delta}\right)}\epsilon_{\rm ff,0}\frac{M_{\rm tot}}{t_{\rm ff}} \tag{52}$$

$$= \frac{1}{(1+\delta)\,\Gamma\left(1+\frac{1}{1+\delta}\right)\Gamma\left(1+\frac{\delta}{1+\delta}\right)}\overline{\epsilon}_{\rm ff}\frac{M_{\rm tot}}{t_{\rm ff}}, \tag{53}$$

where in the second step we have made use of equation (33) to rewrite the SFR in terms of the mass-averaged star formation efficiency $\overline{\epsilon}_{\rm ff}$. We have already noted that the constraints on δ arising from stellar age distributions are inconsistent with those derived from YSO counts, but the total SFRs in both cases are similar. Consulting Tables 3 and 4, we see that YSO counts give $\overline{\epsilon}_{\rm ff} \approx 0.01$ and $\delta \approx 0$, so overall we obtain SFR $\approx 0.01 M_{\rm tot}/t_{\rm ff}$. Stellar ages give $\overline{\epsilon}_{\rm ff} \approx 0.05$ and $\delta \approx 3$, which again gives SFR $\approx 0.01 M_{\rm tot}/t_{\rm ff}$. Thus the global SFR predicted by our best-fitting values of the IE model are roughly the same as those obtained in the ST, CB, or CBD models, and is consistent with the global star formation budget of the Milky Way.

Next consider the GC and GCD models. In this case, evaluation of the integral in equation (50) gives

$${\rm SFR} = f_{\rm GCD}\left[\left(\frac{1+\tau_{\rm coll}}{\tau_{\rm coll}}\right)\epsilon_{\rm ff}\right]\frac{M_{\rm tot}}{t_{\rm ff,0}} = f_{\rm GCD}\left[\frac{\xi}{2}\left(\frac{1+\tau_{\rm coll}}{1+\eta}\right)\right]\frac{M_{\rm tot}}{t_{\rm ff,0}}, \tag{54}$$

where

$$f_{\rm GCD} = \frac{(1+\tau_{\rm coll}\phi_{\rm d})\left[1-\frac{\phi_{\rm d}-1}{\phi_{\rm d}}(1-x_{\rm fb})^{\tau_{\rm coll}}\right]}{1+\tau_{\rm coll}\phi_{\rm d}\left[1-\frac{\phi_{\rm d}-1}{\phi_{\rm d}}(1-x_{\rm fb})^{\tau_{\rm coll}+1}\right]} \tag{55}$$

is the factor by which the SFR is lower in a GCD model than in a GC one due to the extra dispersal at late times; this factor is unity for the GC model, and it also approaches unity for $x_{\rm fb} \rightarrow 1$ or $\phi_{\rm d} \rightarrow 1$, in which limits the GCD models reduces to the GC one. The term in square brackets in equation (54), which we have written in two equivalent ways in order to illustrate the limiting behaviour for $\tau_{\rm coll} \ll 1$ and $\gg 1$, can be thought of as the 'effective' $\epsilon_{\rm ff}$ of the

model. If $\tau_{\rm coll} \gg 1$, i.e. clouds collapse slowly compared to their star formation time-scale, then clearly this term just approaches $\epsilon_{\rm ff}$, equation (54) approaches equation (43), and this model approaches the behaviour of ST, CB, or CBD. If, on the other hand, $\tau_{\rm coll} \ll 1$ so that clouds collapse quickly compared to their star formation time-scale, then the term in square brackets approaches $(\xi/2)/(1+\eta)$. This is just the product of the collapse time measured in units of the free-fall time, $(\xi/2)$, and the fraction of the mass converted to stars rather then lost to the wind, $1/(1+\eta)$. The instantaneous value of $\epsilon_{\rm ff}$ does not matter in this limit, because all the stars form in the final plunge to infinite density.

We cannot apply this result to the Galaxy as a whole, because both the free-fall time $t_{\rm ff,0}$ and the total mass $M_{\rm tot}$ at the start of collapse are unknown – ATLASGAL tells us only the instantaneous mass of clumps whose density is high enough for them to be included in the catalogue, i.e. those for which $t_{\rm ff} \lesssim 0.3$ Myr. However, we can still apply our model just to the ATLASGAL clumps, simply by interpreting the birth rate $\dot{\mathcal{N}}$ as the rate at which clouds become dense enough to be visible to ATLASGAL. Since in the GC and GCD models $t_{\rm ff}$ is monotonically decreasing, we can in this case simply adopt $t_{\rm ff,0} = 0.3$ Myr and set $M_{\rm gas}$ to the total mass of the ATLASGAL samples, and then use equation (54) to compute the contribution to the Galactic SFR provided by those clumps that are dense and massive enough to fall into the ATLASGAL catalogue. This provides a lower limit on the total Galactic SFR.

Inserting the observed mass and free-fall time of the ATLASGAL clumps, we therefore find that the GC and GCD models predict that they should yield an SFR

$$\mathrm{SFR} = 11\,\mathrm{M_\odot\,yr^{-1}}\, f_{\rm GCD} \left(\frac{\epsilon_{\rm ff}}{0.01}\right) \left(\frac{\tau_{\rm coll}}{0.03}\right)^{-1} \left(\frac{M_{\rm tot}}{10^7\,\mathrm{M_\odot}}\right) \left(\frac{t_{\rm ff}}{0.3\,\mathrm{Myr^{-1}}}\right), \quad (56)$$

where we have normalized to our best-fitting value of $\tau_{\rm coll}$ based on observed stellar age distributions, and our numerical evaluation assumes $\tau_{\rm coll} \ll 1$. We can immediately see that there is a serious problem with the star formation budget in the GC model: for the best-fitting parameters arising from stellar age distributions and YSO counts, the observed ATLASGAL clumps should form stars at nearly five times the total SFR of the Galaxy as a whole. The problem becomes even more severe if we recall that ATLASGAL clumps are much denser than the mean density of observed star clusters, and thus likely represent only a small subset of the total star formation in the Galaxy, i.e. $\mathrm{SFR} \gg \mathrm{SFR}(<t_{\rm ff})$.

The GCD model has the potential to perform better, since for it $f_{\rm GCD} < 1$, i.e. the SFR is potentially lower due to the final dispersal phase in this model. We can address this possibility both analytically and numerically. Analytically, note that equations (30) and (55) together imply that

$$f_{\rm GCD} \geq (1+\eta)\epsilon_*, \quad (57)$$

so that values of $f_{\rm GCD} \ll 1$ also imply values of $\epsilon_* \ll 1$, in which case it is difficult to see how bound clusters could form. Indeed, using equation (54), we have

$$\mathrm{SFR} \geq \xi\epsilon_*(1+t_{\rm coll})\left(\frac{M_{\rm tot}}{t_{\rm ff,0}}\right) \gtrsim 16.5\epsilon_*\,\mathrm{M_\odot\,yr^{-1}}, \quad (58)$$

where in the numerical evaluation we have set $\xi \geq 1$, since, as noted above, when one uses the spherical equivalent density (as has been done for the ATLASGAL sample), this inequality holds. Thus the observed SFR of the Milky way is only consistent with a GCD model in which $\epsilon_* \lesssim 0.1$, in which case we expect that almost none of the ATLASGAL clumps could go on form a bound cluster. This seems problematic, since if ATLASGAL clumps cannot go on to form bound clusters, it is unclear what structures can.

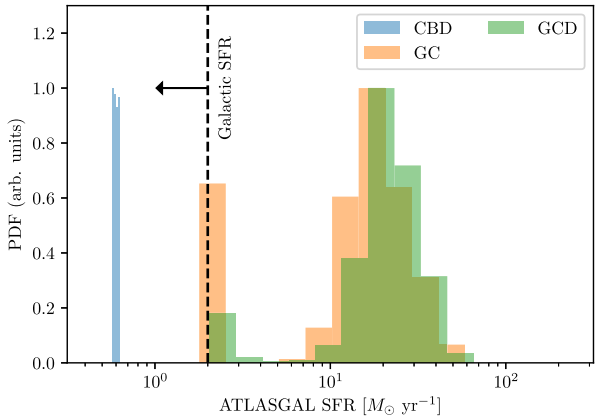

Figure 8. Histograms of predicted SFRs for the gas clumps in the ATLASGAL catalogue, using the CBD, GC, and GCD models with parameters constrained by fitting to the stellar age distribution in young clusters and the number of YSOs per unit gas mass in the ATLASGAL sample. All histograms have been normalized to have a maximum of unity for ease of comparison. The vertical dashed line marks the total SFR of the Milky Way (Chomiuk & Povich 2011); values to the left of this line, indicated by the arrow, are consistent with the total Galactic SFR, while values to the right of it are inconsistent.

We can also use our MCMC analysis address the value of $f_{\rm GCD}$, and whether it allows one to simultaneously match the ATLASGAL and YSO count data. To do so, we proceed as follows. First, since we have seen from Fig. 5 that ATLASGAL YSO counts essentially constrain only $\epsilon_{\rm ff}$, while age distributions constrain other variables but not $\epsilon_{\rm ff}$, we select from our MCMC chains for our fit to the stellar age distribution all samples for which $\epsilon_{\rm ff}$ lies within the 16th to 84th percentile range allowed by our analysis of YSO counts.[5] This gives us a set of parameter values that are consistent with both sets of observations. Secondly, for each sample we compute the quantity $f_{\rm GCD}[(1+\tau_{\rm coll})/\tau_{\rm coll}]\epsilon_{\rm ff}$ (c.f. equation 54), and the corresponding predicted value of the SFR for that set of parameters. The result is a set of predicted SFRs for the ATLASGAL clumps, considering only those parameter values that are also consistent with the data on YSO clumps and age distributions.

We plot the distribution of predicted SFRs in Fig. 8. For comparison, we also plot the corresponding distributions for the GC model (which uses an identical procedure except that $f_{\rm GCD} = 1$ for all samples) and for the CBD model (for which we derive the SFR from equation 43). As expected based on the arguments above and on equations (43) and (56), the CBD model predicts that ATLASGAL clumps form stars at a few tenths of a Solar mass per year, consistent with all of the bound star formation in the Galaxy occurring in them, and perhaps a small amount of unbound as well. The GC model overproduces the SFR of the Galaxy by a factor of $\sim$10. The figure also shows that the GCD model does not do any better than the GC model at matching the observed Galactic SFR; the extra dispersal at the end, once we constrain the parameters that describe it by the observed age distributions and YSO counts, does not allow a significantly lower total SFR for GCD than for GC.

[5]We use the fits to NGC 6530 for this purpose, but the results for the ONC are qualitatively the same. Similarly, using values of $\epsilon_{\rm ff}$ constrained to lie in the 5th to 95th or the 1st to 99th percentile range also does not change the qualitative result.

There is a small tail of parameter space that allows the ATLASGAL sample to have an SFR comparable to that of the entire Galaxy, but even this solution is problematic, since these models are viable only to the extent that one is willing to assume that star formation in the Galaxy occurs exclusively in clumps as dense or denser than the ONC, i.e. the lower density regions like Perseus, Taurus, Ophiuchus, etc., make zero contribution to the Galactic SFR.

The fundamental problem for the GC and GCD models is completely analogous to the one noted by Zuckerman & Evans (1974) for CO-detected molecular clouds, and by Krumholz & Tan (2007) for HCN-detected ones: the model assumes that order unity of the mass in the ATLASGAL clumps will be converted to stars on a time-scale comparable to the free-fall time, which yields an SFR much higher than the one we actually observe in the Milky Way. However, there as an important extra feature here, which is not present in the earlier works. One can avoid the problem of overproducing stars from the CO and HCN data by assuming a very high-mass loading factor, either at all times (in GC) or at late times (in GCD). However, we can now see that this solution is in strong tension with the combined YSO counts and stellar age data. The YSO counts require that the SFR per free-fall time stay nearly constant, so the only way for star formation to accelerate, as required by the observed age distributions, is for the total density in the star-forming gas to rise. For the acceleration to be enough to match the observations, this density increase must occur substantially faster than the gas is depleted by star formation or feedback – in terms of the parameters of our models, we require $\tau_{\rm coll} \ll 1$. However, if the density is increasing much faster than gas is removed by feedback, this in turn implies a high total star formation efficiency. There is no way to simultaneously satisfy the constraints of low SFR per free-fall time in individual clumps and accelerating star formation without also overproducing the total SFR of the Galaxy.

4 SUMMARY AND CONCLUSION

In this paper we investigate a number of candidate scenarios for the formation of bound star clusters, focusing on questions of how the mass is assembled, how it evolves, and how efficiently it forms stars. We do so taking advantage of two significant observational advances over the past few years. The first is the availability of spectroscopically estimated ages for a reasonably complete sample of stars that can be assigned with high confidence to young clusters using *Gaia* kinematics (Kounkel et al. 2018; Prisinzano et al. 2019). These data now show in multiple clusters that star formation in clusters is an accelerating but extended process, i.e. the SFR increases over time, but the total duration of star formation is several free-fall times, so that $\sim$ 30–50 per cent of the stars in any given cluster are more than three free-fall times old, and $\sim$ 5–10 per cent are as old as ten free-fall times. Explaining this accelerating but extended star formation history requires a model in which either the total mass of gas available for star formation increases with time, the efficiency of star formation at fixed gas mass and density increases with time, or the mean density increases with time, leading to an increase in the SFR – these scenarios roughly correspond to the models of conveyor belt star formation, IE of star formation, and global hierarchical collapse that have previously appeared in the literature.

The second data set of which we make use is a large sample of star-forming gas clumps from the ATLASGAL survey (Schuller et al. 2009; Csengeri et al. 2014; Heyer et al. 2016). that are well-matched to young star clusters in terms of mass and density, but which are still very gas rich and thus likely represent a slightly earlier evolutionary state. Such gas clumps show a very tight correlation between the mass of gas, its mean density, and the number of YSOs embedded within it, which together constrain the rate at which the gas produces YSOs. We carry out a Bayesian forward-modelling treatment of the observational uncertainties and possible biases in these data set, including the effects of selecting only gas-dominated systems and of changes in the gas properties on time-scales shorter than the YSO lifetime, and we find that these factors do not significantly alter the overall constraint on how efficiently gas produces YSOs. The tight correlation of gas properties with YSO counts rules out the possibility that the star formation efficiency per free-fall time is time-dependent, ruling out models where the observed acceleration of star formation is due to a time-dependent increase in star formation efficiency per unit mass per unit free-fall time.

We finally consider the global star formation budget of the Milky Way, and show that the scenarios of global hierarchical collapse and conveyor belt star formation predict that the observed ATLASGAL clump population will yield very different total rates of star formation in the Galaxy. The collapse scenario is only able to recover the observed acceleration of star formation if clumps collapse globally on a time-scale shorter than that on which they initially form stars locally, since otherwise depletion of the gas by star formation yields a star formation history that decelerates rather than accelerating. However, the requirement for global collapse to occur before a significant fraction of the mass can form stars in turn requires that the ATLASGAL clumps produce stars at a rate that exceeds the entire SFR of the Milky Way, let alone the substantially lower rate at which bound star clusters form.

By contrast, the conveyor belt model, first proposed by Longmore et al. (2014), encounters no such difficulties, because it attributes the acceleration of star formation to the fact that gas clumps form stars and accrete simultaneously, so that the gas mass available for star formation tends to increase with time until the gas is dispersed by feedback. We further find that accretion at a rate that varies with time as $\dot{M}_{\rm acc} \propto t^3$, as generically predicted for the gravitational collapse of mass reservoirs with fixed bounding pressure (Goldbaum et al. 2011), produces a distribution of stellar ages consistent with that observed in young clusters. We therefore conclude that the best available explanation for all of the available observational constraints is that bound star clusters form in a conveyor belt mode, where gas accretes at an increasing rate, but the central cluster-forming region is not in a state of global collapse, and has a star formation efficiency per unit mass that is both low and roughly constant in time.

ACKNOWLEDGEMENTS

We thank E. Vazquez-Semadeni, J. Ballesteros-Paredes, A. Palau, G. C. Gomez, and M. Zamora-Aviles for comments on the manuscript, and we thank the anonymous referee for a helpful report. MRK acknowledges support from the Alexander von Humboldt Foundation, and funding from the Australian Research Council through the Future Fellowship (FT180100375) and Discovery Projects (DP190101258) funding schemes. CFM acknowledges support by National Aeronautics and Space Administration (NASA) through NASA ATP grant NNX13AB84G. This research made use of ASTROPY,[6] a community-developed core PYTHON package for Astronomy (Astropy Collaboration 2013, 2018).

[6] http://www.astropy.org

REFERENCES

Adamo A., Kruijssen J. M. D., Bastian N., Silva-Villa E., Ryon J., 2015, MNRAS, 452, 246
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Azimlu M., Martínez-Galarza J. R., Muench A. A., 2015, AJ, 150, 95
Barnes A. T. et al., 2019, MNRAS, 486, 283
Beccari G. et al., 2017, A&A, 604, A22
Caldwell S., Chang P., 2018, MNRAS, 474, 4818
Chandar R., Fall S. M., Whitmore B. C., Mulia A. J., 2017, ApJ, 849, 128
Chomiuk L., Povich M. S., 2011, AJ, 142, 197
Csengeri T. et al., 2014, A&A, 565, A75
Da Rio N., Tan J. C., Jaehnig K., 2014, ApJ, 795, 55
Da Rio N. et al., 2016, ApJ, 818, 59
Dekel A., Krumholz M. R., 2013, MNRAS, 432, 455
Elmegreen B. G., 2000, ApJ, 530, 277
Evans N. J. et al., 2009, ApJS, 181, 321
Evans N. J., II, Heiderman A., Vutisalchavakul N., 2014, ApJ, 782, 114
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Fűrész G., Hartmann L. W., Megeath S. T., Szentgyorgyi A. H., Hamden E. T., 2008, ApJ, 676, 1109
Ginsburg A., Bressert E., Bally J., Battersby C., 2012, ApJ, 758, L29
Goldbaum N. J., Krumholz M. R., Matzner C. D., McKee C. F., 2011, ApJ, 738, 101
Gutermuth R. A., Megeath S. T., Myers P. C., Allen L. E., Pipher J. L., Fazio G. G., 2009, ApJS, 184, 18
Gutermuth R. A., Pipher J. L., Megeath S. T., Myers P. C., Allen L. E., Allen T. S., 2011, ApJ, 739, 84
Heyer M., Gutermuth R., Urquhart J. S., Csengeri T., Wienen M., Leurini S., Menten K., Wyrowski F., 2016, A&A, 588, A29
Hillenbrand L. A., Hartmann L. W., 1998, ApJ, 492, 540
Huff E. M., Stahler S. W., 2006, ApJ, 644, 355
Jaehnig K. O., Da Rio N., Tan J. C., 2015, ApJ, 798, 126
Jeffries R. D., 2007, MNRAS, 381, 1169
Jeffries R. D., 2017, Mem. Soc. Astron. Italiana, 88, 637
Johnson L. C. et al., 2016, ApJ, 827, 33
Kharchenko N. V., Piskunov A. E., Schilbach E., Röser S., Scholz R.-D., 2013, A&A, 558, A53
Kim D., Lu J. R., Konopacky Q., Chu L., Toller E., Anderson J., Theissen C. A., Morris M. R., 2019, AJ, 157, 109
King I. R., 1962, AJ, 67, 471
Klessen R. S., Burkert A., 2000, ApJS, 128, 287
Kounkel M. et al., 2018, AJ, 156, 84
Kroupa P., Aarseth S., Hurley J., 2001, MNRAS, 321, 699
Kruijssen J. M. D., 2012, MNRAS, 426, 3008
Kruijssen J. M. D., Dale J. E., Longmore S. N., 2015, MNRAS, 447, 1059
Kruijssen J. M. D. et al., 2019, Nature, 569, 519
Krumholz M. R., Tan J. C., 2007, ApJ, 654, 304
Krumholz M. R., Dekel A., McKee C. F., 2012, ApJ, 745, 69
Krumholz M. R., McKee C. F., Bland-Hawthorn J., 2019, ARA&A, 57, 227
Kuhn M. A., Hillenbrand L. A., Sills A., Feigelson E. D., Getman K. V., 2019, ApJ, 870, 32
Kuznetsova A., Hartmann L., Ballesteros-Paredes J., 2015, ApJ, 815, 27
Kuznetsova A., Hartmann L., Ballesteros-Paredes J., 2018, MNRAS, 473, 2372
Lada C. J., Lada E. A., 2003, ARA&A, 41, 57
Lada C. J., Lombardi M., Roman-Zuniga C., Forbrich J., Alves J. F., 2013, ApJ, 778, 133
Lee Y.-N., Hennebelle P., 2016a, A&A, 591, A30
Lee Y.-N., Hennebelle P., 2016b, A&A, 591, A31
Lee E. J., Chang P., Murray N., 2015, ApJ, 800, 49
Lee E. J., Miville-Deschênes M.-A., Murray N. W., 2016, ApJ, 833, 229
Longmore S. N. et al., 2013, MNRAS, 429, 987
Longmore S. N. et al., 2014, Protostars and Planets VI. University of Arizona Press, Tucson, AZ USA. p. 291
Matzner C. D., Jumper P. H., 2015, ApJ, 815, 68
Messa M. et al., 2018, MNRAS, 473, 996
Motte F., Bontemps S., Louvet F., 2018, ARA&A, 56, 41
Murray N., Chang P., 2015, ApJ, 804, 44
Ochsendorf B. B., Meixner M., Roman-Duval J., Rahman M., Evans N. J., II, 2017, ApJ, 841, 109
Palla F., Stahler S. W., 2000, ApJ, 540, 255
Preibisch T., 2012, Res. Astron. Astrophys., 12, 1
Prisinzano L. et al., 2019, A&A, 623, A159
Rathborne J. M. et al., 2014, ApJ, 786, 140
Reggiani M., Robberto M., Da Rio N., Meyer M. R., Soderblom D. R., Ricci L., 2011, A&A, 534, A83
Ryon J. E. et al., 2014, AJ, 148, 33
Schuller F. et al., 2009, A&A, 504, 415
Soderblom D. R., Hillenbrand L. A., Jeffries R. D., Mamajek E. E., Naylor T., 2014, Protostars and Planets VI. University of Arizona Press, Tucson, AZ USA. p. 219
Tan J. C., Krumholz M. R., McKee C. F., 2006, ApJ, 641, L121
Toalá J. A., Vázquez-Semadeni E., Gómez G. C., 2012, ApJ, 744, 190
Tobin J. J., Hartmann L., Furesz G., Mateo M., Megeath S. T., 2009, ApJ, 697, 1103
Urquhart J. S. et al., 2018, MNRAS, 473, 1059
Vázquez-Semadeni E., González-Samaniego A., Colín P., 2017, MNRAS, 467, 1313
Vázquez-Semadeni E., Palau A., Ballesteros-Paredes J., Gómez G. C., Zamora-Avilés M., 2019, MNRAS, 490, 3061
Vutisalchavakul N., Evans N. J., II, Heyer M., 2016, ApJ, 831, 73
Walker D. L., Longmore S. N., Bastian N., Kruijssen J. M. D., Rathborne J. M., Galván-Madrid R., Liu H. B., 2016, MNRAS, 457, 4536
Ward J. L., Kruijssen J. M. D., 2018, MNRAS, 475, 5659
Zamora-Avilés M., Vázquez-Semadeni E., 2014, ApJ, 793, 84
Zamora-Avilés M., Vázquez-Semadeni E., Colín P., 2012, ApJ, 751, 77
Zuckerman B., Evans N. J., 1974, ApJ, 192, L149

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

submit_supp.pdf
Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TEX/LATEX file prepared by the author.