We address the referee's criticisms in reverse order of thier report.

The referee's only concrete criticism of our manuscript was about a particular sentence in our abstract, which begins "[t]he class of metastable initial data forms a horizon over longer time scales than suggested by the lowest order perturbation theory..." We have therefore clarified this statement by adding the phrase "at computationally accessible amplitudes," which we have already emphasized throughout the main text of the manuscript (so we have made no other changes).

The referee also claims that it is "arbitrary" to classify initial data as unstable or metastable based on the scaling of t_H for t_H larger than some chosen t_{fit}, specifically because the choice of t_{fit} affects the classification. However, if we are interested in studying whether a given set of initial data has perturbative $t_H \sim \epsilon^{-2}$ scaling at finite amplitude, we have to choose which amplitudes to fit, since large enough amplitudes should not be expected to have perturbative scaling and would therefore ruin a fit. This does not make it an arbitrary choice but rather a reasoned choice as explained in our previous response. Furthermore, we note that all the initial data classified by our definition as metastable lie on the boundary of the island of stability, strongly suggesting a real physical meaning — it takes longer for initial data there to enter the regime to perturbativity (as the referee agrees). The only question then is whether to highlight the "less perturbative" initial data with a separate classification, which, as we have made clear in our manuscript and responses, applies in the computationally accessible regime. Since the point should clearly be highlighted in some manner, any argument over the language used is therefore a matter of semantics. While the referee does not like making a separate class, we believe it is entirely appropriate. In any case, a disagreement over semantics should not prevent publication of a manuscript when authors and referee agree on the science, as we do in this case.

Finally (or initially), the referee's major criticism of our manuscript is that we fit t_H with either the function $a\epsilon^{-p} + b$ or $a\epsilon^{-p} + b + c\epsilon^2$. According to the referee, this is too "restrictive," and we should include odd powers in the Laurent series as well, for example fitting to $t_H = a\epsilon^{-p} + b\epsilon^{-1} + \cdots$. As the referee is aware, if not from the wider literature on AdS gravitational collapse, then certainly from our manuscript and previous responses, such an expansion would be physically nonsensical for the system we study. Specifically, the perturbative expansion of any quantity is well-known to contain only alternating powers of ϵ, and t_H has even powers. To be very clear on this point, we demonstrate that the leading terms of the perturbative expansion of t_H are $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(1)$ below:

By definition, the metric variables $A(t,x) = e^{\delta(t,x)} = 1$ at lowest order, ie, these are the values for empty AdS.[1] Then, the scalar field variables $f = \{\phi, \Phi, \Pi\}$ are all $\mathcal{O}(\epsilon)$, or $f(t,x) = \epsilon f_1(t,x) + \cdots$ (take ϕ as $\mathcal{O}(\epsilon)$ by definition and find the lowest order contribution to Φ, Π from equations (2,3) in our

[1] Strictly speaking, $\delta(t,x) = \delta_0(t)$ also describes empty AdS, but a time reparameterization always allows us to set $\delta_0 = 0$, and the literature universally does so for convenience.

manuscript). From equations (4-6) in the manuscript, any correction to A, δ is quadratic in the fields, so $A = 1 + \epsilon^2 A_2(t,x) + \cdots$, $\delta(t,x) = \epsilon^2 \delta_2(t,x) + \cdots$. Returning to (2,3) again, we see that the only possible corrections from the metric variables are $\mathcal{O}(\epsilon^3)$, so we can write $f(t,x) = \epsilon f_1(t,x) + \epsilon^3 f_3(t,x) + \cdots$ — since they are not sourced by the gravitational interaction, any $\mathcal{O}(\epsilon^2)$ terms can be removed by a renormalization of ϵ. Plugging back into equations (4-6), we see that the next correction must be $A = 1 + \epsilon^2 A_2(t,x) + \epsilon^4 A_4(t,x) + \cdots$, $\delta(t,x) = \epsilon^2 \delta_2(t,x) + \epsilon^4 \delta_4(t,x) + \cdots$. This is sufficient for our needs, but generalization to arbitrary order by induction proceeds similarly.

Now consider a perturbative expansion of t_H, defined as the earliest such time that $A(t_H, x_H) \leq A_{thresh}$ for some point x_H and some small fixed threshhold value A_{thresh} (as described in our manuscript). In other words, $\ln A$ changes by an order 1 amount between $t = 0$ and t_H. The time evolution of A is determined by an additional constraint from the Einstein equations, see eqn (30c) from our reference [7], which we can write as

$$\frac{d \ln A}{dt} = -2 \sin(x) \cos(x) A e^{-\delta} \Phi \Pi \ .$$

Clearly, the lowest order contribution to the time derivative is $\mathcal{O}(\epsilon^2)$, leading to the leading scaling $t_H \sim \epsilon^{-2}$. At this lowest order in perturbation theory, we expect that each function $A_i(t_H, x) \sim \epsilon^{-2}$ with $1 + \epsilon^2 A_2(t_H, x_H) = A_{thresh}$ (expanding the log gives $\dot{A}_2 \sim 1$ for example). Including the next order of perturbation theory, $\epsilon^4 A_4 \sim \epsilon^2$, so $A(t_H, x_H) - A_{thresh}$ is now $\mathcal{O}(\epsilon^2)$ using the lowest order t_H. This difference is corrected by time evolution by long enough to correct A_2 by an order unity amount, which should only take a time of order unity. Thus, the first subleading correction to t_H should be order 1. It would indeed be surprising if $\dot{A}_2 \sim 1$ for $\mathcal{O}(\epsilon^{-2})$ time before "stalling" and reducing to $\dot{A}_2 \sim \epsilon$ precisely at the lowest-order value of t_H (which is necessary for the first correction to t_H to be $\mathcal{O}(\epsilon^{-1})$ as the referee suggests). The possible loophole is that the horizon location x_H also shifts with each successive order of perturbation theory. This can only decrease t_H at each order compared to assuming x_H fixed. However, A_2 at the two putative x_H values must then be within $\mathcal{O}(\epsilon^2)$ of each other to be re-ordered by the A_4 term. This is again incompatible with one reaching A_{thresh} at $\mathcal{O}(\epsilon^{-1})$ time before the other, closing the loophole.

As we see, finding a nonzero ϵ^{-1} in the Laurent expansion of t_H would run counter to any type of perturbation theory. So, contrary to the referee's suggestion, finding a better fit to the $t_H \sim \epsilon^{-2}$ scaling by including an ϵ^{-1} term would *not* in fact indicate perturbative instability but rather that the initial data in question are strongly *nonperturbative* in that amplitude range.

The referee's criticisms of our manuscript are trivial (and corrected), spurious, and demonstrably incorrect, as described above. We therefore expect that our manuscript is ready for publication with no need for further review.