Editor Comments

While the comments from the first reviewer are quite sparse and can likely be addressed with the addition of only a few extra references tying the work to the broader literature and perhaps a few more sentences in the Discussion, the second reviewer raised a number of deeper issues. I ask that you please work to address the second reviewer's first and second comments especially, though I found all of the second reviewer's comments to be valuable.

Based on the reaction of both reviewers and my own reading, I also ask you to consider adding a statement of current limits in scope and interpretation (qualified as being a necessary next step in the sophistication of such models) to the Introduction (perhaps as an extra sentence or two in the last paragraph, starting on line 51?).

We have added a few sentences to the end of the Introduction (line 58), where we discuss the scope and limitations of our model.

I ask that you revise your manuscript according to these recommendations after which I expect to solicit a second opinion from the reviewers, but I have every expectation that you will be able to satisfy the concerns raised and that the end result will be a compelling and important contribution to the literature that can then be found acceptable for publication.

Thank you. We found the reviewer's comments to be very helpful, and have addressed them below. The corresponding edits to the manuscript are in blue font, which we believe improve the manuscript greatly.

Reviewer # 1

Comments:

I liked reading this manuscript, and found it acceptable as written. However, I also had many reactions that were along the line of wanting a more complete assessment of the implications of this study, and how specifically it goes beyond any previous similar investigation.

Thank you for your helpful comments. To our knowledge, there is only one other study that has constructed a model of coupled social and climate dynamics (Beckage et al. 2018). We detail how our study goes beyond theirs in the first paragraph of the Discussion section (line 275).

For example, people are organized socially in ways that go from individual actions to families/neighborhoods, to cities/states, and interacting countries. The particular scale at which this study would be relevant is thus unclear. It seems to be based on individual social norms, but many of the implications are at global or international scales. Some assessment of this aspect would be useful, and could even greatly extend the manuscript or analysis if the journal is able to take a longer manuscript. It could be characterized as further evaluation of the effect of size of the social unit to be considered.

Great point, thank you. The model does indeed construct social dynamics at the scale of the individual (citizen), which is reflected in the payoff functions to mitigate / not mitigate. A natural extension to the model would be to consider not only the behaviour of individuals, but that of larger bodies such as corporations / governments, which presumably would require different payoff functions. However, we consider this extension to be beyond the scope of this current paper. We have appended the Discussion section (line 303) with a more detailed discussion of social scale.

Reviewer # 2

Comments:

This paper is a novel attempt to link a simplified social model to a simplified global earth system model that has at least similar qualitative behavior to very complex models that project global temperature range over the next 200 years or so. This paper could well motivate a much more complex push towards incorporating social system models for human behavior into global climate models and this is clearly something that has been lacking in the IPCC scenarios to date, The simplicity of the model is both a blessing and a curse - it is sufficiently simple to pick out broad qualitative behaviors for projections as affected by various model components, but it could be argued that it is far too simple and thus provides a non-useful "cartoon".

We agree that making quantitative projections of socio-climate trajectories requires a model that captures more complexity than our proposed model. However, we see our model as a tool to generate mechanistic insights regarding how different social processes can qualitatively impact the socio-climate system, which is better investigated with a simple model that is easy to understand and yet captures what we believe to be the key social features. We also hope that this simple model will motivate further extensions and more complex models to investigate similar ideas.

I provide some comments that might assist the authors in enhancing the manuscript to make it a bit less readily criticized for being too simple.

Thank you, your comments have improved the manuscript considerably and we describe how we addressed them in the following paragraphs.

1. There is not really any discussion of model evaluation or what criteria are set prior to model development to ensure the resulting model results provide new insight to meet the model objectives. For example, it could be argued that a realistic model for linking social behavior to climate should have the possibility that the model leads to continued increasing GHG emissions not ameliorated by behavior and not assume a priori that mitigating behavior will spread. Yet this model seems constructed with the assumption that mitigation behavior will spread, though the timing for this depends in a somewhat complicated manner on model parameters. Can the model as constructed with the chosen range of parameters lead to increasing GHG emissions similar to those observed in many IPCC scenarios (e.g. 5 degree increase by 2100, or over longer time periods)? It doesn't appear so from the simulations run, which implies the model is constructed specifically to meet the objectives of fostering mitigation behavior.

Good point, we did not make explicit the criteria that were set for the model, for it to achieve our objectives for the paper (stated in line 53). As such, have appended the Introduction to include four criteria that were required for the model to meet our objectives (line 58). We feel that this greatly clarifies our adopted approach to modelling socio-climate dynamics, with its advantages and limitations.

We also agree that socio-climate models should not assume a priori that mitigative behaviour will spread throughout the population. Although not shown in the original manuscript, there are parameter configurations for our model where mitigative behaviour does not spread. In particular, for a sufficiently large strength of social norms and cost to adopt mitigative practices (within the parameter bounds), the strategy of non-mitigative behaviour is dominant throughout the time-frame considered (up to 2200). (We note that social norms in our model reinforce majority behaviour and can therefore support either mitigative or mitigative behaviour.) We have included a simulation of this 'worst-case' scenario the SI Appendix (Figure 3). Under this parameter configuration, the temperature dynamics are then consistent with that of the RCP 8.5 scenario (4 degrees by 2100).

2. There is an inherent asymmetry in the model structure that is assumed and not explained. The utility for mitigators has a constant cost while that for non-mitigators is a function of the temperature anomaly. It is not clear why both should not be a function of the temperature anomaly - why should the utility to mitigators of the cost of mitigation not depend on temperature anomaly? It is a perceived cost associated with the temperature analogy so the assumption is either that the perceived cost for the anomaly for mitigators doesn't depend on the size of the anomaly, or that the anomaly cost is small relative to the cost, beta, that is assumed. If it is a simplifying assumption then it needs to be justified. However the cost structure for the temperature anomaly is generally much higher than the value of beta.

This is a good point. As it turns out, an alternative formulation that provides temperature dependence for the payoff to mitigate yields the same set of equations, but we agree that the way we derived it was confusing. The updated manuscript uses an this alternative formulation (Methods, line 93).

In summary, mitigators now receive a payoff

$$e_M = -\alpha + cf(T) + \delta x.$$

which is the same as before, except now there is a temperature specific term providing a positive utility to mitigate that increases as the temperature increases. (The parameter representing mitigation costs has also been relabelled from beta to alpha - we now use beta to represent a ***net*** *cost to mitigate). We use the sigmoidal form for the incentive to mitigate based on temperature anomaly, as we assume that this utility is proportional to the negative costs that an individual would incur if they did not mitigate, as coded in the payoff for* e_N*:*

$$e_N = -\gamma - f(T) + \delta(1 - x).$$

This payoff now includes an additional parameter (gamma), that imposes a cost on non-mitigators such as a carbon tax. When the payoff functions are combined, the parameters alpha and gamma are combined to form beta, the net cost to mitigate. Constructing the social dynamics as in Methods, we arrive at

$$\frac{dx}{dt} = \kappa x(1-x)(-\beta + (1+c)f(T) + \delta(1-x)).$$

This equation yields the same dynamics as before, just with the different parameter labelling $(1+c)f_{\max} \longleftrightarrow f_{\max}$*, i.e. the parameter* c *gets incorporated in* $f_{\max}$ *. Therefore this alternate model formulation (which we have now adopted) does not change our results.*

What if the cost for non-mitigators was a constant? This means omega = 0 in eq. 8, but that case is not considered (according to the value range chosen for omega in the supplement table). Then, depending upon the relative costs for mitigators and non-mitigators, there are model equilibria (in the social model eq 6) of x = 0, 1 and a possible intermediate one. Initial conditions will determine which is approached in this situation - though the forcing function epsilon in the other model component may mean such an equilibria is not reached.

We did not consider the system dynamics for $\omega = 0$*, as a fundamental assumption of the model is that the social system depends upon the climate system. In the case* $\omega = 0$ *the social system evolves independently, but is an interesting system in its own right. Upon doing a stability analysis, we find that there is indeed an intermediate equilibrium of this system*

$$x^* = \frac{1}{2\delta}\left(\beta - \frac{f_{\max}}{2} + \delta\right),$$

but when it is physically relevant (in [0,1]), it is unstable, and marks the point that separates two basins of attraction containing the stable states of full mitigation and full non-mitigation . The fate of the climate system would then be in the hands of where the initial condition for x lies! I suppose we omitted the possibility of w=0, as the paper goes by the assumption that experiencing or forecasting an increase in global temperature is a key driver for mitigative action.

What matters in all this is the value of beta relative to the values of f(T) but beta has a mean value of 1 and max of 1.5 and the f(T) function has a peak of 5 and values for midpoint of 2.5 (Fig 3 in supplement). This forces the model generally to have higher utility for mitigators leading to an increase in mitigation just due to the parameter values chosen. Why is the mid value for the perceived costs in f(T) taken so much higher than the value of beta has at max? What is the impact of making these the same?

Good point - the balance between beta and f(T) is important in determining whether mitigative behaviour begins to spread. More specifically, if the social system is near the state x=0, then mitigative behaviour will only begin to spread once

$$f(T) > \beta + \delta,$$

i.e. the cost due to global temperature rise must exceed the cost to adopt mitigative practises and the cost to go against the current norm of non-mitigation. With the baseline parameter values used in the manuscript ($\beta = 1, \delta = 1, f_{\max} = 5, \omega = 3$*) this transition begins to take place for* $T \approx 2.4$ *(by eyeballing where f(T)=2). In answer to your question, if we construct f(T) to have a mid value similar to beta (by taking* $f_{\max} \approx 2$*), then the payoff to not-mitigate will always be higher around* $x = 0$ *and mitigation will never spread, no matter how high T gets.*

We also consider it reasonable to assume that f(T) will be much larger than beta after a few degrees of warming, given much of the economics literature that argues the costs of global warming will far exceed the costs required to prevent it. We have included this argument in Methods (line 168) with a citation to the Stern review on the economics of climate change who argue this case.

3. There is a fixed initial condition for fraction of mitigators which is set very low. Since there is considerable effort made to consider changes in other parameters and sensitivity of results, this should be done for this initial condition too. At the least there should be some discussion of the impact of a population of mitigators that is initially much lower (say x=.01 or rather higher - say x=.1) on model dynamics. For example, I assume that if x initially is set smaller there would be an even longer time period in which mitigation fraction drops which could lead to much longer time horizons for mitigation behavior to spread.

Thanks, this is a good point. We have now included x0 in the sensitivity analysis with lower and upper bounds 0.01 and 0.1 respectively. SI Appendix Table 2 and the tornado plot in Figure 4 have been updated to reflect this. Interestingly, the peak temperature anomaly is relatively insensitive to x0, when compared to the parameters governing the strength of the social processes, like norms and learning. As you suggested, a smaller initial proportion of mitigators results the social system taking longer to transition to mitigative behaviour, and therefore resulting in a larger peak temperature anomaly. We now mention this insensitivity in the manuscript (line 245).

4. There is no discussion of a striking result of the sensitivity runs in Fig 4 - the asymmetry in sensitivities that appears in many parameters. Presumably this arises from some inherent nonlinearities in the model and if so, why are some impacts much more asymmetric than others - can this be explained?

The asymmetry in the sensitivities is certainly an interesting feature, thanks for pointing this out. We have appended the Results section (line 252) to highlight this feature and what we can learn from it.

5. Eq 7 assumes a linear extrapolation of temperature based on the past. Is there any evidence for this at all from the variety of survey results on attitudes they mention? Is it just a simplifying assumption? Since the model uses non-linear functions for the key model component f(T) why not consider this non-linear as well?

Assuming that humans linearly extrapolate from past temperatures is certainly a simplifying assumption, but serves as a starting point for more complex models to elaborate on. To our knowledge there is no evidence as to the functional form that this should take. We consider linear extrapolation to be a reasonable choice, since many climate forecasts show a roughly linear trajectory for the temperature over the next 100 years (SI Appendix, Figure 1). We bring this point into the revised paper (Discussion, line 314).

6. What time horizon parameter is used in Fig 2 - is it tf and if so what is tp? if it is tf then the range of values used for this figure is much larger than that in the supplement table.

Thanks for prompting clarification here. The 'time horizon' in Figure 2 refers to t_f, the number of years with which an individual projects the temperature into the future. The caption for Figure 2 has been updated to reflect this. The lower and upper bound for t_f have been adjusted to 0 and 50 years respectively, and this has been made consistent with the range time-horizon parameter values shown in Figure 2. All sensitivity analyses have been rerun with the updated parameter bounds on t_f and x_0. The only noticeable change is the sensitivity bar in Figure 4 for t_f, which is now larger, reflecting the drastic increase in peak temperature anomaly when there is zero foresight (t_f =0) in the population.

7. Figure 3 in the text has beta going up to 2 but in the table in the supplement it says the upper bound used was 1.5

We have adjusted the upper and lower bounds of the axes in Figure 3 to be consistent with the upper and lower bounds used for the sensitivity analyses (which can be found in Table 2 SI Appendix).

8. minor error - line 96 "begin" should be "being"

Thank you, this has been updated.