1 Response to Comments From Reviewer 1

1.1 I really feel given the nature of this paper (it is not the definitive guide to heat and health in Houston) that it needs to be fully capable of reproducibility. I therefore don't feel the part of the algorithm set out in section 2.4 is sufficient. For example I had to read to find details for α_p etc., some of the other comments (the posterior is fully known) were a little glib. I think the entire sampler could be set out in the text, but if not an appendix could be used. Links to code would also be helpful (I appreciate there are confidentiality issues with some of the data).

Valid point. We rearranged the paper to put all of the information relevant to the algorithm in 2.4 together in the same section, making it easier to find details about how the model was fit. We also provided additional description to clarify what was meant by our comments such as "the posterior is fully known."

1.2 I don't think the comments on convergence are adequate. On page 13 line 10; how does MCSE alone provide evidence of convergence? I think more is needed.

We have attached traceplots to this document and can include them in the supplementary materials for the paper if desired. Additionally, we added further detail about how MCSE provides evidence of convergence on page 14 with the following statement:

> In order to assess convergence, we examined traceplots for each of the parameters and used Monte Carlo standard error (MCSE). Jones et. al. (2006) calculate MCSE by dividing the chains into batches and using each batch mean to calculate variance between batches. Small MCSE values occur only when the variance between batches is low (i.e., each batch is approximately equal), thereby providing evidence of convergence.

1.3 I wasn't sure about claiming IG(2,1) or the Dirichlet prior were non-informative. Using population counts sounds very informative (or was it the expected values that were used).

In response to this comment we added a significant amount of justification in Section 2.3 detailing exactly why these priors are non-informative. Additionally, we clarified our comments in order to avoid the miscommunication that we were using population counts in our prior. The population counts only come into play through the data in the likelihood.

1.4 It would be useful to understand some of the covariate posterior information; does the 0.51 versus 0.49 for males versus females section 3.3) really mean anything? This might be a great parameter to set out the trace plots and density plots which would help provide information on convergence as well as understanding the relevance of what sounds like a modest difference to me.

For this, they ask for trace plots and density plots, but I'm not sure how that answers the question he asks. I've included the posterior delta densities, so that if desired you can mentally multiply that image by the proportions to ballpark how many people that difference might represent, but I could actually create a figure of the distribution of the counts and include it in the text if you think it's a good idea.

1.5 Why do a spatial misalignment (Brunsdon and Comber)? There are comparable numbers of census tracts and grid squares; using census tracts directly sounds like a tractable GIS problem (the denominator is the area of the census tract, points-in-polygons can count the necessary points). I couldn't see anything in the mathematics of the intensity that would prevent the census tracts being used, and you remove a layer of approximation.

We responded to this comment by adding the following text to the article:

> A potential solution to avoid realigning the census data to the grid cells is to let the census block groups equal the $\mathcal{G}_k$ denoted in (**??**) rather than the grid cells. However, the heat information used in this research (described in detail in Section **??**) is provided at a 1-km^2 resolution and these grid cells were chosen to align exactly with the available heat data. Using the block groups in place of the grid cells would necessitate aligning the heat information to the census block group areas. Because heat is our primary variable of interest we prefer realigning the demographic information to the grid associated with the heat data as opposed to the alternative. Therefore, we proceed having aligned the population data as described.

1.6 Heat enters as an upper level covariate, with ecological fallacy risks? Are the census data not available to study how tsome of these features vary with key other demographics? I appreciate all the census data are disclosure limited but some multiway information on air conditining and age band could really limit the potential problems with confounders in the aggregate analysis.

In order to address this concern, we examined the impact of additional demographic covariates from the census data on our model. We also included all two-way interaction terms between the various covariates as potential predictors in our model in order to mitigate the potential problems with confounders (see Table X), as suggested by the reviewer.

1.7 Although you make a comment in the end about the lack of suitable model fit diagnostics there is a lot more that is already available. WAIC / Gneiting and Tillmans's proper scoring rules seem appropriate given a major outcome is a risk map.

The comment in our article was not about model fit diagnostics that allow you to compare different models, but rather about diagnostics that are an objective measure of how well the chosen model fits the data. Neither Tillman Gneiting's proper scoring rules or WAIC provide a solution to this particular problem. Can you add a discussion of the proper scoring rules here? Additionally, Gelman et. al. (2014; *Statistics and Computing* argue that "a cost of using WAIC is that it relies on a partition of the data into n pieces, which is not so easy to do in some structured-data settings such as time series, spatial, and network data" and suggest that this makes WAIC an inappropriate measure for the type of spatial model we use in this paper. However, following the work of Leininger, we did implement some simple posterior predictive checks in order to show that our analysis is generally congruent with the observed data.

1.8 From an applications point of view I'd really like to see the counts (posterior predictive density) as well as the probabilities in figure 4.

I've added in the delta posterior densities, giving an idea of the counts in general, but should I go ahead and add the counts for the ages specifically as well?

2 Response to Comments from Reviewer 2

2.1 The authors need to explain explicitly how to combine/connect and use data information from 2 different periods of time (heat-related emergency calls during 2006-2010 vs. mortality data between 1999 and 2006).

Via email, you suggested that we show that the demographic makeup of Houston didn't change much over the window that we consider. Do you have any idea how to show that? Should I examine census data over that period? What do we do in the case where there was a major demographic shift?

2.2 The prior assumptions and choices need to be justified especially for those non-informative and vague priors. Also, "Because the priors for α_p, γ_p" etc., "are all conjugate, their posterior distributions are known..." is not clear to me that the claim is valid. It would also be helpful if the results from the MCMC are presented so to confirm convergence.

We added additional detail justifying our prior assumptions and also explicitly demonstrated the conjugacy of the Dirichlet prior. Because we have so many parameters, we prefer not to include trace plots in the paper, but we have attached them to this document and can add them in the supplementary materials for the paper if desired.

2.3 The labels of colours in figures 2 and 3 are hard to see and the 95%CI in figure 3 is not clear.

We have improved the visibility of the labels, and made the widths of the 95% confidence intervals more obvious.