Response to Comments From Reviewer 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 $\boldsymbol{\alpha}_p$ 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).*

   We strongly welcome comments to make our work more reproducible. 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." Jacob: Can you are least include a reference to exactly where the changes you made here occur in the paper. That'll just make it easier on the second review.

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.

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.

4. *It would be useful to understand some of the covariate posterior information; does the 0.51 versus 0.49 for males versus females 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.*

   We have done two things to address this comment: First, we provided highest posterior density (HPD) intervals for the probabilities in Table 4, showing that there is no overlap for the male and female mortality probabilities and suggesting that the difference is statistically significant. We feel that in terms of understanding uncertainty this interval is a more useful summary than trace plots, and that a density plot does not add enough additional information beyond the HPD interval to justify its inclusion. Second, in order to provide additional context for this difference we have included density plots for the δ parameters for the 911 calls and mortalities. Recall that δ represents the expected total number of incidents for each health outcome across the spatial region. Multiplying this number (≈ 15228 for mortalities) by the probabilities associated with each sex gives an estimate for the size of the difference (i.e., about 7766 incidents of male mortality vs 7461 for females). Recognizing that this small difference in probabilities represents around 400 deaths suggests that this effect is practically significant as well, though not to the degree of some other factors, such as race/ethnicity.

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.

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.

   Jacob: I am not sure I follow the Table 2 in the paper comparing all the different models. Did each of these models have the same DIC?

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. For example, continuous rank probability scores are to assess the predictive ability of a model rather than the fit of a model to available data. 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.

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.*

   In order to provide information about the counts succinctly, we included the density plots for δ_{911} and δ_{Mort} in Figure 4. You can get an estimate of the number of mortalities and 911 calls associated with each age by multiplying the probabilities in Figure 4 by the appropriate δ value. Including the information about δ allows us to provide information about the counts associated not just with each age, but also with the race/ethnicity probabilities and gender probabilities in Tables 4 and 5.

Response to Comments from Reviewer 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).*

   In the paper, we added text that provides a more lucid discussion of the assumptions needed to combine the two relative risk surfaces. The paper now reads:

The primary assumption we make in (5) is that the relative risk surfaces for each health outcome share a common, underlying risk surface that is constant with respect to time. Naturally, the assumption would be violated if there is temporal variability between the relative risk surface in 1999-2006 (the time window for the mortality data) and 2006-2010 (the time window for the 911 call data). Unfortunately, given the data available to us, our assumption of a time-invariant relative risk surface is not able to be evaluated. That is, we would need information on the relative risk of mortality between 2006-2010 to investigate any inherent differences in the surface between 1999-2006 and 2006-2010 (which we don't have). Likewise, we would need 911 call data for the years 1999-2006 to assess any temporal differences in the risk of a 911 call (which we also don't have). While the possibility of a time-varying risk surface would be interesting from an epidemiological perspective, we leave this for future work.

An additional implicit assumption with (5) is the independence assumption among the log relative risks for each health outcome given the overall, "cumulative" log relative risk surface $\boldsymbol{\mu}$. For the application considered here this assumption is likely to, approximately, hold because the time frame for the 911 call data is 2006-2010 while the time frame for the mortality data is 1999-2006 (a minimal overlap); however, in other applications independence may not hold. In such cases, a multivariate spatial model (see Genton and Kleiber (2015) for a discussion) could be used in place of (5). We note the possible benefit of modeling dependence between the log relative risks but we do not explore this option here.

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.

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.