CANSSI - Letter of Submission

Reviewer 1

This reviewer asks no specific questions. Is there anything to respond to?

Reviewer 2

1. What sort of research would be needed to construct the new bcAQHI once the new bcGAIM is built and computationally implemented?

   With a Poisson response distribution and the response variable being the appropriate health outcome, the bcGAIM will output a relative risk for every combination of (measured or forecasted) pollutant values input into the model. The bcAQHI will take the relative risks as inputs and output warnings, based on cutpoints for the relative risk. The key task is determining the cutpoints, which we may do in a model-based way or using expert knowledge. Finally, forecasted pollutant values for next-day predictions will be provided by Environment Canada.

2. The cGAIM itself is by no means a new idea. The original model for a single group (K=1) goes back to Hardle (1993). In fact, one seems to get Hardle's model if one drops the fk's from the model. Wang et al (2015) presents a multigroup version (K>1) to get around the curse of dimensionality, the whole point of this approach. But special cases were published between 1993 and 2015.

   There are new features in the cGAIM – it considers constraints and groupwise additive index terms, while much of the existing literature only considers one or the other. While Hardle, Hall, and Ichimura (1993) examine a single index model and Wang et al. (2015) consider a multiple index model, neither consider constraints. Two papers that consider constrained estimation are Xia and Tong (2006), where the authors constrain s to be monotonic and the components of α to be non-decreasing, and Fawzi et al. (2016), where the authors constrain the components of α to be non-negative and sum to one but do not constrain s. In comparison, the cGAIM allows any linear constraint to be placed on α and different shape constraints on s including montonicity, convexity, and concavity (Masselot et al, 2020). Additional comparisons between the cGAIM, bcGAIM, and models in the literature are given in the *Research Aims* section.

3. The third main topic seems the most novel in as much as it will show how the new bcAQHI might be used to assess COVID-19 mortality. Of course, it would seem more reasonable to me to build a new bcAQHI designed specifically for that purpose. And that led me to wonder about the health outcome to be used to fit the bcGAIM-the all-causes mortality-to get the alpha and in turn the index. For example, ozone would seem more relevant as a risk index for asthma and PM2.5 for COPD. Why not just publish the pollutant concentrations themselves?

   The bcGAIM estimates the relative risk of an observed air pollutant mixture, which is transformed into the bcAQHI by identifying cutpoints for the relative risk. For the COVID-19 model, we would like to report relative risks so we would fit the bcGAIM using COVID-19 mortality to estimate the relative risk of the pollutant mixture. We would also use COVID-19 specific fixed effects and confounders for this model. We can apply the bcGAIM in this way to any study, such as an epidemiological study, for which we would like to report relative risks. Finally, if desired, we could apply cutpoints to the estimated relative risks to obtain warning levels for use in a COVID-19 specific air quality index.

4. The conversion of the cGIAM to a bcGIAM should be feasible. A major challenge will be the big data problem since daily health counts for all-cause mortality will be modelled, and this using a semi-parametric model. But the task would be simplified by the intended approach of designing a different model for each Canadian city, how many we don't know. But they do not intend to incorporate random city effects in the Bayesian framework to enable strength to be borrowed deficiency in the intended approach, but perhaps a compromise needed for feasibility.

   There are significant challenges relating to the size of the data. We will be examining 25+ regions in Canada over a 20+ year period (>6,000 days) of daily data. Fitting the bcGAIM to each city independently is computationally feasible, while fitting a hierarchical model across 25 regions is less feasible but perhaps more desirable. We have given further thought to this goal since submitting the LOI, and will be implementing various Laplace approximations in the target density to lessen the computational burden. This would allow us to produce estimates of the relative risks of the pollutant mixture across the 25+ regions. The details of the proposed approximation are discussed in the *Research Aims* section.

5. A complicated issue and I had to do a lot of digging to figure out how the team of Investigators was assembled. In part, this involves those listed as Collaborators. The LOI could have done a better job of clarifying the links of the Investigators and Collaborators in the proposed project.

   There are four investigators in the project. Patrick Brown and Meredith Franklin are at the University of Toronto, while Cindy Feng is at Dalhousie University in Halifax. A collaborator, Daniel Rainham, is also at Dalhousie University. The fourth investigator, Fateh Chebana, is at INRS in Quebec. Patrick has relations with Health Canada and is currently researching the effects of air pollution, while Fateh has relations with INSPQ and has (with Pierre Masselot) developed the cGAIM. Cindy Feng and Daniel Rainham...

   There is one postdoctorate, one PhD, and one Master's student involved in the project. The postdoctorate student will split time between the University of Toronto and Dalhousie University. When in Toronto, his activities will be supervised by Patrick and Meredith. When in Halifax, his activities will be s...upervised by Cindy and Daniel. One graduate student will be based at the ... and co-supervised by Patrick and Meredith. The other will be based at ..., and co-supervised by .... The graduate students will also spend summer sessions at the home institutions of the other investigators. Finally, the undergraduate summer students will be under the supervision of the investigator that applied for their funding. The *Anticipated Roles of Trainees* and *Anticipated Organization of Collaboration* sections contain more details on how collaboration is organized, and the role of the investigators.

6. It is challenging to coordinate and run such a program successfully and we don't get a clear impression from the LOI that the applicants have thought about this issue very much. What is clear is that it is designed to provide the pipeline from data through to the AQHI. That is excellent. What is not clear is how the collaboration is to be managed and what I for one would like to see if a proposal is invited is an active group interaction plan.

   Regarding organization, there will be weekly meetings between supervisors and trainees. Trainees that are working on more than one project may attend more than one weekly meeting, particularly if they are working with off-site collobartors. The postdoctoral student will organize a monthly meeting that everyone attends, which will be held virtually. Each trainee will be supervised by one or more of the investigators, and the postdoctoral student will split time between locations to facilitate knowledge transfer and to help build team cohesion.

7. The discussion of the bcGIAM is well done including the parts about the implementation of the computation strategies. But the LOI should have explained it better. For a start the big picture with its three main objectives should have been better described in general terms and how they relate to one another before diving into details re the cGIAM.

   The bcGAIM project has three research objects: to develop a multi-pollutant air quality health index (AQHI), to develop multi-pollutant exposure models for various health effects, and to investigate how mixtures of pollutants affects daily COVID-19 mortality. These three tasks are all applications of the bcGAIM. The bcAQHI is derived from the relative risks estimated by the bcGAIM, the exposure models are bcGAIM models (with different fixed effects, smooth functions, and mixtures), and the COVID-19 model is the bcGAIM with daily COVID-19 mortality as the outcome. Thus, the methodological innovations are found in the bcGAIM, namely in developing interpretable priors and implementing approximate inference algorithms. The research aims - the bcAQHI, epidemiological studies, and COVID-19 studies - are potential high-impact applications of the bcGAIM that we will be pursuing.

Reviewer 3

1. The naïve case fatality rate that the investigators plan to explore with regard to Objective 3 is subject to errors caused by an undercount of both the numerator and the denominator and known to be a poor measure of the mortality risk of the disease.

   While reported COVID-19 mortality rates are subject to reporting error, and there may be inaccuracies in the reported data released by government officials, it is the best data available for the COVID-19 inquiry. If there are anomolies such as persistent under-reporting, under-reporting on weekends or holidays, different inclusion criteria (long-term care cases and deaths may not be reported), or any others we can address them during data processing or model-fitting. For instance, we can add a day-of-the-week effect to account for delayed reporting on weekends, have the model impute missing values, or build a more complex model that uses COVID-19 mortality counts and non-COVID-19 mortality counts to infer excess deaths attributable to COVID-19.

2. Although four methodological advancements of the bcGAIM are listed in the Methods section, those seem to be rather incremental changes of the cGAIM.

   The *Research Aims* section has additional material explaining the differences between the cGAIM, bcGAIM, and existing models.

3. How the bcGAIM helps with developing a simple, intuitive air quality index that simultaneously accounts for the health effects of multiple air pollutants is not explained.

   This was clarified in the response to the first question by Reviewer 2.

4. The potential for impact in statistics and inferential data science seems to be marginal considering that methodological advancements of the bcGAIM listed in the Methods section are deemed to be rather incremental changes of the cGAIM.

   The bcGAIM has novel innovations that will enable new statistical reasoning to applied to observational data that involves mixtures, such as the multi-pollutant problem. There are a number of statistical challenges in developing the bcGAIM. For one, fitting a model to daily observations across 25+ regions in Canada is a significant computational task. The bcGAIM attempts to address this by applying a (non-linear) one-dimensional function to a linear combination of related covariates, which both eases the computational burden and improves the intrepretability of the model. To accomplish this, we must set shape-constrained priors on s and estimate α. These are both challenging problems and are discussed in more detail in the *Research Aims* section. It also has additional material to clarify the differences between the cGAIM, bcGAIM, and existing models in the literature.

Reviewer 4

1. The team has the potential to provide an excellent environment for interdisciplinary training of students. It would be helpful to add the names of the faculty supervisors/collaborators to the mentoring plan. The part for the roles of personnel could be further clarified.

   Clarification can be found in the answer to the second question posed by Reviewer 2, as well as the *Anticipated Roles of Trainees* and *Anticipated Organization of Collaboration* sections. These two sections contain more details on how collaboration is organized, and the role of the investigators.

Reviewer 5

1. A potential weakness would be that while well motivated by multi-pollutant modeling, the investigators could strengthen the proposal by identifying other applications where this form of model would be applicable.

   The *Research Aims* section contains a brief discussion of other applications, with some further discussion in *Anticipated Roles of Trainees*.

2. What is the relationship between the linear combination that goes into the smooth function and the air quality index? How will the index provide measures that indicate it is safe or not based on the smooth function and or linear combination of the exposures? In terms of data, is this based on hospital admissions (asthma or other conditions or only mortality). More details on data sources related to aims would be helpful.

   The bcGAIM estimates the relative risk of the linear combination of the pollutants, and is translated into an AQHI based on cutpoints on the levels of the relative risk. For the AQHI, we are collaborating with epidemiologists and they will advise on the appropriate health outcome to use for the AQHI.

3. While the researchers describe this as a constrained or shape constrained model, the proposal lacks details about what shape constraints are desired. There is a significant literature on Bayesian shape constrained modeling (monotonicity in particular) but limited references are provided. How does the proposed research build on this and what will be novel and contribute to general statistical methodology? Is this a straightforward model to fit in STAN or code directly or are there methodological advancements to be made there?

   The *Research Aims* section contains additional details on the bcGAIM, and more extensive references to the literature on shape-constrained inference. There are a number of challenges to fitting the bcGAIM in Stan. **TODO**: Ask Patrick.

4. The researchers cite Stringer et al (2020) as developing a Bayesian single pollutant version of a case-crossover model using non-MCMC methods such as Integrated Nested Laplace Approximations (INLA). Given the space limitations of the proposal it is not clear that such results will immediately carry over to the Bayesian shape constrained model where both the (constrained) smooth function of the linear combination of the exposures, the weights in the linear combination of exposures and the smooth functions of confounders has to be estimated, in addition to the other smooth functions of confounders. INLA like methods have been used to solve an array of complex problems, so this may be feasible. While this might provide computational efficiencies in point estimation, how does this address the importance of uncertainty quantification of the Bayesian model over the frequentist model of Masselot et al (2020)?

   The *Research Aims* section explains the INLA-like approximation in more detail. To summarize those additions, the bcGAIM has link function $g(\lambda_t) = X^T\beta + s(\alpha^T Z_t) + f_1(W_{1,t}) + \ldots + f_K(W_{K,t})$. Conditional on α, we can simplify the estimation problem by considering parameters ϕ, θ, and α and estimating $\pi(\eta|Y,\theta,\alpha)$, $\pi(\alpha|Y,\theta)$, $\pi(\theta|Y)$, and $\pi(\eta|Y) = \int \pi(\eta|Y,\theta,\alpha)\pi(\theta|Y,\alpha)\pi(\alpha|Y)d\theta d\alpha$ (the last one numerically). The Laplace approximation can be applied to $\pi(\eta|Y,\theta,\alpha)$ and $\pi(\theta|Y)$ (within Stan), and $\pi(\theta|Y,\alpha)$ can be estimated using HMC. We will also develop an approximation of $\pi(\theta|Y,\alpha)$ outside of Stan. We will release both Stan models (exact and approximate), as well as the standalone approximate inference algorithm, in an R package.

5. In discussing priors to induce shape constraints, the proposal rejects the idea of placing a prior on the expansion of the smooth function, but rather to place it directly on s. What types of priors on functions spaces are going to be used; Gaussian Process priors or others? How are the constraints incorporated?

   We will be considering shape constraints on Gaussian processes – they are flexible function classes and are widely used. The shape constraints would be incorporated by re-parameterizing the Gaussian process, or exploiting its structure to enforce the desired shape constraint. Ideally, a re-paramterization would create a parameter whose value directly relates to the shape constraint. For example, Kamal et. al. (2020) proposes a re-parameterization of the anisotropic Matern function where one parameter controls

the anisotropic ratio of the Matern random field, and another the anisotropic angle. Alternatively, the Guassian process contains mathematical structure that allows us to introduce an object that plays a similar role. The *Research Aims* section contains a brief discussion of shape-constrained Bayesian inference, and additional discussion on our approach to developing shape-constrained priors.

6. While the impact of pollutant exposure and COVID-19 exposure is suggested by Wu et al (2020), how does this model relate to the models for total mortality? Will this be a joint model for COVID versus non COVID deaths with common smooth s function or different smoothed functions? Or will there be different linear combinations of exposures or the same? Or will this utilize the proposed Air-Quality index? What data are available and do they provide the necessary information about potential confounders or other covariates? (socio-economic status, access to health insurance, housing status (group living such as nursing homes, dorms, single family, number of family members etc) co-morbidities, etc. Are individual level data available or is this aggregated data at say a county level? Missing data is clearly an issue with COVID-19 deaths but is not addressed.

   <!– Wu et. al. (2020) fit a model with a negative binomial response distribution and COVID-19 mortality aggregated at the country-level as the response variable. They use the national long-term average PM2.5 level as their exposure, and incorporate 20 potential confounding variables (age, population density, number of individuals tested, socioeconomic variables, smoking, obesity, etc.) into their model. They found that a 1 $\frac{\mu g}{m^3}$ increase in PM2.5 exposure was associated with an 8

   To investigate the relationship between air pollution and COVID-19 mortality, we will fit the bcGAIM to COVID-19 mortality. The granularity of COVID-19 case and mortality reporting varies widely depending on the reporting region. In Canada, the Government of Ontario's Treasury Board Secretariat provides many COVID-19 data sets, including "Long-Term Care Home COVID-19 Data", "Confirmed positive cases of COVID-19 in Ontario", and "Status of COVID-19 cases in Ontario". The "Confirmed positive cases of COVID-19 in Ontario" data set contains age, gender, location (by public health unit), and the patient outcome. The "Status of COVID-19 cases in Ontario" data set contains daily tests completed, test outcomes, case outcomes, current hospitalizations, and current patients in ICUs.

   It is important to note that none of these data sets were available at the time of our LOI submission. The Government of Ontario has gradually made additional information available to the public, a trend that will likely continue in Ontario and other reporting regions. It's very reasonable to expect more data in Ontario (and elsewhere) to be made available over time. Already, we have access to rich COVID-19 data sets for Ontario. We can incorporate demographic and socio-economic variables at the census tract or city-level, which is similar to the county-level aggregation seen in Wu et. al. (2020). We could, for example, compare COVID-19 mortality in low-income vs. high-income regions, in majority white vs. majority minority regions, or in densely populated vs. less densely populated regions.

   The John Hopkins COVID-19 Github page lists their data sources. In the United States, reporting is done by state agencies and occasionally by large cities such as New York and Illionois. The finest resolution these regions report data most often aggregated at the county-level, and variables such as age, gender, and race/ethnicity are usually provided. Many regions also provide hospitalization and ICU data. We can again introduce demographic or socio-economic variables at the country-level. While there are still data limitations, fairly rich COVID-19 data sets are available today that will allow us to conduct meaningful inquiries into the relationship between COVID-19 and air pollution exposure.

   Ideally, we would be able to fit a model that estimates both COVID-19 and non-COVID-19 mortality, perhaps with the goal of estimating excess deaths attributable to COVID-19. However, while daily COVID-19 mortality data is being made publicly available, other daily mortality data is not pubicly available and recent Canadian will take time to access. Therefore, this extension to the COVID-19 model is likely to be undertaken in the 2^{nd} or 3^{rd} year of this project. Finally, note that such a model would involve using both sets of data sets described in next response.

7. As this is a very application motivated proposal it would be useful to know what data are available for each of the aims and how they will be integrated.

   The daily air pollution data is from the National Air Pollution Surveillance (NAPS) Program, a network

of 250 stations across Canada that is managed by Environment Canada and Environment and Climate Change Canada. Potential environment confounders for the air pollution model, such as temperature and humidity, is from Environment and Climate Change Canada. The COVID-19 data is from official data sources provided by provincial authorities. These data sets are all publicly available. The daily health outcome data is provided courtesy of Health Canada and INSPQ (the Quebec Public Health Institute), and is not publicly available.

The bcGAIM is a general model - it contains fixed effects β, smooth functions of potential confounders $f = (f_1, \ldots, f_k)$, and a smooth function of a linear combination s whose coefficients α are also of interest. In our applications, the response is a mortality or morbidity outcome of interest. We will be fitting two versions of the bcGAIM to these two problems - one for the multi-pollutant model that produces the bcAQHI, and one for the COVID-19 investigation. The response for the COVID-19 model will be COVID-19 mortality, and it will also have different fixed effects and confounders. More details on data sources related to the COVID-19 model are provided in the response to the previous question.

The daily health outcome data provided by Health Canada is aggregated count data at the census tract or city level. Depending on the province in question, the COVID-19 case and mortality data has variables such as age, socio-economic status, and so on. Since the multi-pollutant model and COVID-19 models are being developed separately, the data sets identified for each model will be combined for use in each model. It is less likely they will be combined across models, as discussed in the previous response.

8. The proposal mentions developing random effects in the smooth function. What do these capture and what is the motivation? i.e spatial random effects, treating the weights as random effects (to allow spatial variation). Additional clarity would be helpful.

   By random effects, we mean random walks. **TODO**: Ask Patrick.