Title of the Project: Statistical Methods for Daily Mortality and Multiple Environmental Risk Factors

Name and affiliations of lead investigators

- Patrick E. Brown
  - Centre for Global Health Research, St. Michael's Hospital
  - Department of Statistical Sciences, University of Toronto
  - pbrown.ca
- Fateh Chebana
  - Centre Eau Terre Environnement, Institut national de la recherche scientifique, Québec.
  - inrs.ca/en/research/professors/fateh-chebana
- Cindy Feng
  - Department of Community Health and Epidemiology, Dalhousie University
  - medicine.dal.ca/departments/department-sites/community-health/our-people/our-faculty/cindy-feng.html
- Meredith Franklin
  - Keck School of Medicine, University of Southern California (until July 2021)
  - School for the Environment and Department of Statistical Sciences, University of Toronto (from July 2021)
  - keck.usc.edu/faculty-search/meredith-franklin

List of collaborators, titles, and affiliations

Name: Kamal Rai
Title: PhD student
Affiliations: Centre for Global Health Research, St. Michael's Hospital
Department of Statistical Sciences, University of Toronto

Name: Daniel Rainham
Title: Professor
Affiliations: School of Health and Human Performance, Dalhousie University
Healthy Populations Institute, Dalhousie University

Name: Hwashin Shin
Title: Scientist
Affiliation: Environmental Health Science and Research Bureau, Health Canada

Name: Prabhat Jha
Title: Director
Affiliation: Centre for Global Health Research, St. Michael's Hospital

Name: Céline Campagna
Title: Responsable scientifique
Affiliation: Équipe Changements climatiques et santé, Institut National de Santé Publique du Québec

Name: Pierre Masselot

Title: Research Fellow
Affiliation: London School of Hygiene & Tropical Medicine, United Kingdom

List of partner organizations

- The **Centre for Global Health Research, St. Michael's Hospital** will lead the health sciences research component of the project, providing data from the US and India and time of research staff to work on manuscripts. The Toronto-based component of the team will be located at the Centre for Global Health Research (CGHR) and will be integrated into the Geospatial Mortality research group Dr. Brown leads. A portion of the Toronto PhD student's salary will be funded though a CGHR research grant.
- **L'Institut National de Santé Publique du Québec** has the mandate to analyse, monitor and evaluate determinants of health, including environmental pollutants, and will give access to the Quebec provincial health databases. In-kind support will also include health expertise for the statistical design and interpretation of research, including co-supervision of students.
- Hwashin Shin at **Health Canada** is central to this project. Her needs for an improved air quality indicator instigated the discussions which lead to this application. Health Canada will provide detailed cause-specific daily mortality and morbidity counts. Dr. Shin has been funding Dr. Brown's team through Health Canada research contracts, and intends to continue doing so throughout this project.

Research Aims

Overview

There is growing interest in developing a simple, intuitive air quality index that combines multiple pollutants while simultaneously estimating their health effects (Dominici et al. 2010; Stieb et al. 2008; Bopp et al. 2018). Typically, health effects associated with exposure to air pollution focus on one pollutant at a time, but this ignores the fact that realistiacly, exposures are complex mixtures of multiple pollutants (Dominici et al. 2010). Thus, an air quality index that reflects this complex mixture is needed and should account for the levels and relative contributions of multiple air pollutants together. In this proposal we present an improved statistical method that allows us to model multiple environmental pollutants while simultaneously estimating their health effects. With this method we are also able to conduct inference on the health effect estimates. We demonstrate our approach by focusing on quantifying the short-term health effects associated with air pollution mixtures at the population level.

The *constrained groupwise additive index model* (cGAIM), introduced by Masselot et al. (2020), is a vehicle for providing a multi-pollutant health index. Given λ_{it} representing a particular health outcome for an individual i on day t, the cGAIM is

$$\lambda_{it} = \exp\left[X_{it}^T\beta + s(\alpha^T Z_{it}) + f_1(W_{1it}) + ... + f_K(W_{Kit})\right],$$

where β parameters are the fixed effects of the potentially time-varying linear covariates X_{it}, and $f_1, ..., f_K$ are smooth functions that account for potentially non-linear covariates or confounders W_{kit} $(k = 1 ... K)$. The distinguishing feature of cGAIM is the smooth function s whose argument is a linear combination of variables Z_t, for example fine particulate matter ($PM_{2.5}$) and ozone (O_3). The α parameter is a vector of estimated weights on the entries of Z_t, and gives their relative contributions. The smooth functions s and the f_k might be composed of spline functions or Gaussian processes such as random walks.

To date, cGaim has been developed for Poisson models, namely where the outcome is case counts (e.g. number of asthma-related hospital visits in a county). We will expand its usage to case crossover models, which have seen increased attention in the air pollution literature (Wei et al. 2019; Stringer et al. 2020a). In case crossover models we can examine individual-level data, whereby the exposure on an event or case day is compared to the exposure on referent or *control days* for each case. For example, the air polution concentrations on a day of an individual's hospital visit will be conditioned on the same individual's exposure on pre-selected control days, perhaps chosen as the air pollution concentrations on same day of the week from the previous two weeks. Case crossover models use a partial likelihood for the probability the event occurs on the case day rather than the control days. The advantage of case crossover models is that individual-level confounders are automatically adjusted for, as are risk factors which vary slowly or not at all, or are the same on the case and control days. The challenge introduced by case crossover models is the likelihood depends on non-linear combinations of the latent variables (i.e. s).

Estimating α is the main statistical challenge with the cGAIM, which Masselot et al. (2020) accomplish with frequentist inference methods that use sequential quadratic programming.

We will develop a Bayesian methodology for inference with the cGAIM — the bcGAIM — which will estimate and fully quantify the uncertainty around α as well as propagate the uncertainty into inference on s. The bcGAIM will also able to efficiently handle higher dimensional α and Z_{it}, which is a significant improvement over cGAIM and an important feature of the model as we aim to model complex multi-pollutant mixtures.

Outcomes and applications

This project brings together the methodological components of several interdisciplinary and collaborative research activities in which the four investigators have independently been engaged.

The primary driver of this research is the need for an improved air quality warning system, which Health Canada and the Institut national de santé publique du Québec have separately approached Drs Brown and Chebana (respectively) about. Currently the Canadian AQHI is composed of relative risks estimated from cohort studies, and estimated risks for individual pollutants are summed to create a log-relative risk which is in turn converted to a 10-point scale. This is likely to over-estimate risk, Franklin and Schwartz (2008) found that the effect of ozone on non-accidental mortality was "substantially reduced" after adjusting for particle sulfate and Liu et al. (2019) found significant differences in the percentage change of all-cause mortality attributable to $PM_{2.5}$ and PM_{10} after adjusting for NO_2 or SO_2. Furthermore, there is evidence that some health outcomes are nonlinearly related to pollution measurements (Feng et al. 2016). Dr. Brown's group has developed a linear multi-pollutant case/crossover model (Huang et al. 2020) whereas Dr. Chebana's has used a cGAIM with a frequentist time series model (Masselot et al. 2020). The proposed bcGAIM is a natural extension of, and merging of, these two methods.

A second driver of this project is the environmental epidemiology research undertaken by the investigators in collaboration with health science researchers. The Centre for Global Health Research, where Dr. Brown is partly based, has history of producing papers on global mortality in high-impact journals. The Million Deaths Study in India has 13 years worth of cause-specific mortality data geocoded to point locations and with smoking and diet information about the deceased and from healthy respondents. Dr. Franklin has a number of highly cited papers on air quality and mortality in environmental health journals. With our collaborators, including Prabhat Jha in Toronto and Daniel Rainham at Dalhousie, we will use bcGAIM to produce papers for the top-ranked medical journals.

The third motivation for this CRT is the surge in availability of daily mortality data brought on by the COVID-19 pandemic. The relationship between daily COVID-19 deaths and air pollution levels has recently become an active area of research. Wu et al. (2020) find that a 1 μg increase in long-term exposure to ambient $PM_{2.5}$ increases the COVID-19 mortality rate by 15%. We will relate COVID-19 incidence and mortality to air pollution in major urban centres worldwide, where possible focusing on deaths outside long-term care homes.

A key reason the bcGAIM model is ideal for the above problems is it will produce parameters which are interpretable. The α coefficients give the relative importance of each pollutant (at each lag), and $s(\cdot)$ is the relative risk from a basket of exposures. Unsupervised methods such

as principal components analysis and clustering can be difficult to interpret (Davalos et al. 2017). A popular nonparametric method is Bayesian Kernel Machine Regression (BKMR), which models an exposure-response surface via a kernel function (Bobb et al. 2015). Using a hierarchical Bayesian variable selection method, it can select one pollutant from a group of correlated ones, and is interpreted by visualizing cross-sections of a potentially high-dimensional exposure-response surface. The bcGAIM will provide similar flexibility to the BKMR, while being able to meet the communication needs of inter-disciplinary research teams.

Methods

The bcGAIM will make four methodological advancements for modeling health effects of mixtures of exposures. These are:

1. develop bcGAIM, a Bayesian inference methodology for high dimensional cGAIM's in case-crossover models;
2. create an efficient, non-iterative computational algorithm for bcGAIM's based on Laplace approximations;
3. develop non-parametric forms of the dose-response effect which encourage or enforce monotonicity; and
4. engage in interdisciplinary and applied research projects with our subject-area collaborators.

Model

A simple two-pollutant version of bcGAIM has been implemented by Dr. Brown's group (under contract from Health Canada) in the MCMC software Stan (Carpenter et al. 2017) with a Poisson response variable. This will be extended to having an α which is 12- or 15-dimensional, with three pollutants (O_3, $PM_{2.5}$, and NO_2) at time lags of up to five days. Converting the Poisson likelihood to case-crossover, moving to a highly-parallelized cloud platform, and increasing the dimensionality will initiallyrequire only modest amount of additional coding, although the algorithm is unlikely to function properly without a substantial amount of further modification and optimization. This is because we expect that α will not always be well identified, and the results will be sensitive to parametrizations and prior distributions.

The major task in this component of the research will be to find reparametrizations and multivariable prior distributions that enable prior elicitation from subject-area specialists. We will adopt the penalized complexity prior framework of Simpson et al. (2017), specifying a base model with some suitably chosen values α_0 and deriving a prior for α which corresponds to an exponential prior on the KL distance between $s(\alpha Z)$ and $s(\alpha_0 Z)$. Expert advice will be used to set α_0 and the rate of the exponential prior. One possible scenario would be to have α_0 with an entry of 1 for $PM_{2.5}$ at lag 1 and zero otherwise, and an extremely strong exponential prior on the KL distance. This prior would encourage a conventional single-pollutant single-lag model, other pollutants would be inferred to have negligible effect on health outcomes unless the data provided considerable information to the contrary. A weaker prior with

α_0 having multiple non-zero entries would allow for stronger multi-pollutant effects. Box constraints, such as the entries of α being positive, should be relatively simple to implement.

Anticipating that the posterior of α will often have flat ridges, interpreted as overall pollution is known to be harmful but specific components may or may not be, communicating results to health scientists will be challenging. Finding intuitive and interpretable low-dimensional functions of α which convey the 'known' and 'unknown' components of the results will be a second important task in this objective.

Computation

For the second methodological aim, we will develop non-MCMC inference methods similar in spirit to INLA (Rue et al. 2009). Case-crossover models are not compatible with the original INLA methodology, and an INLA-like algorithm for these models was recently developed by Stringer et al. (2020a) and further improved in Stringer et al. (2020b). The Latent Gaussian approximation in INLA separates the parameter space into covariance parameters θ and linear predictors $\eta = (\beta, \theta, f)$, and considers $\pi(\eta|Y,\theta)$, $\pi(\theta|Y)$, and $\pi(\eta|Y) = \int \pi(\eta|Y,\theta)\pi(\theta|Y)d\theta$ (the last one numerically). INLA performs approximate inference on θ by estimating $\phi(\theta|Y,\phi)$ with a normal distribution with mean θ^* and variance Σ^*.

Let us translate this reasoning to the bcGAIM, which has link function $g(\lambda_t) = X^t\beta + s(\alpha^T Z_t) + f_1(W_{1,t}) + ... + f_K(W_{K,t})$. Note that conditional on α, $\alpha^T Z_t$ is known. Thus, 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 $\pi(\eta|Y) = \int \pi(\eta|Y,\theta,\alpha)\pi(\alpha|Y,\theta)\pi(\theta|Y)d\theta d\alpha$ (the last one numerically). The first and third densities in the integrand, $\pi(\eta|Y,\theta,\alpha)$ and $\pi(\theta|Y)$ are well-suited to the Laplace approximation while $\pi(\alpha|Y,\theta)$ can be estimated using HMC. Introducing two Laplace approximations will lessen the computational burden, and enable us to fit a hierarchical bcGAIM model to air pollution data. This will allow us to produce national estimates of air quality while fitting the bcGAIM to over 25 regions across Canada, each with over 6,000 daily observations, an otherwise daunting computational task. Therefore, this non-MCMC inference method will provide significant computational and ease-of-use benefits, and will expand the types of problems and number of users who can use the bcGAIM methodology. To facilitate use by other researchers, all bcGAIM software will be released in an R package.

Monotonicity

The third methodological aim will consider ways of encouraging (or forcing) the relative risk function $s(\cdot)$ to be monotonic. Monotonicity for non-parametric smoothing was formalized by Ramsay (1988), and an early example of monotonic additive index models is Xia and Tong (2006). The cGAIM from Masselot et al. (2020) allows for any linear constraint on α and different shape constraints on s including monotonicity, convexity, and concavity. For frequentist inference monotonicity can be incorporated as a restriction of the parameter space during optimization, whereas Bayesian inference involves integrating out all possible realizations of s (not only the mode). Golchi et al. (2015) state "While a rich literature

exists on monotone function estimation, interpolation of monotone functions with uncertainty quantification remains an understudied topic." For this reason much of the previous literature on additive index models is not relevant to the development of bcGAIM.

Initially we will consider different Gaussian processes for s which are more likely to be monotonic than random walks, such as a first-order random walk plus drift. Reparametrizations of higher-order random walks with suitably chosen priors for the boundary values (i.e. first and last point for RW2's) could also produce random functions with a low probability of having local optima. A more sophisticated approach is offered by Golchi et al. (2015), where sequential constrained Monte Carlo is used to take posterior samples from a process with positivity constraints on the derivatives.

Finally, consider an approach similar in spirit to bcGAIM. Bürkner and Charpentier (2020) propose a Bayesian model to estimate ordinal predictors with monotonic effects. They employ a simplex parameter ζ to model normalized differences between categories, and a scale parameter b. The prior on b expresses prior knowledge on the average differences between adjacent categories, while the prior on ζ expresses prior knowledge on individual differences between adjacent categories. The authors suggest an $N(0, \sigma)$ prior on b and a Dirichlet(α) prior on ζ. Then, σ and α would express how heavily average and individual differences between adjacent categories are penalized. Not only are ζ and b interpretable, but so are the prior parameters σ and α. The bcGAIM seeks to achieve this ease of interpretation of its parameters and priors. This will encourage adoption of the bcGAIM in other research areas, which is one of the goals of this project.

Applications

In a highly cited paper with many eminent co-authors, Liu et al. (2019) estimate short-term effects of $PM_{2.5}$ on mortality in 652 cities and produce a pooled global estimate. All-cause mortality risk is shown to be roughly 4% higher at $100\mu g/m^3$ than at $5\mu g/m^3$, with the curve steeper at low values than higher values. A Poisson time series model with regression splines was used. We will reproduce this analysis in as many cities as possible using bcGAIM, improving on Liu et al. (2019) by using case-crossover models, monotonic semi-parametric models, and accounting for the combined effects of multiple pollutants.

COVID-19 incidence and mortality counts have been made publicly available for many countries, much of the data are available sub-nationally, with age and sex information, and for individuals in and outside of nursing homes. While incidence data are both incomplete and with a level of completeness varying spatio-temporally, better measures of incidence will become available from antibody studies. The Centre for Global Health Research runs the Action to Beat Coronavirus study (abcstudy.ca), which started collecting blood samples from a representative sample of the Canadian population in May 2020. As of October, 8000 samples have been returned and a second round of testing is underway. With repeated samples an estimate of true incidence over time can be obtained and changes in under-reporting inferred. We will use bcGAIM to quantify the effect of air quality on COVID-19 incidence, mortality, and the case fatality rate.

References

Bobb, J. F., L. Valeri, B. Claus Henn, D. C. Christiani, R. O. Wright, M. Mazumdar, J. J. Godleski, and B. A. Coull (2015). "Bayesian Kernel Machine Regression for Estimating the Health Effects of Multi-Pollutant Mixtures". In: *Biostatistics* 16.3, pp. 493–508.

Bopp, S., A. Richarz, A. Worth, E. Berggren, and M. Whelan (2018). "Something from Nothing: Ensuring the Safety of Chemical Mixtures". In: *Ensuring the safety of chemical mixtures, Publications Office of the European Union, EUR* 29258.

Bürkner, P. and E. Charpentier (2020). "Modelling Monotonic Effects of Ordinal Predictors in Bayesian Regression Models". In: *British Journal of Mathematical and Statistical Psychology*.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software* 76.1.

Davalos, A. D., T. J. Luben, A. H. Herring, and J. D. Sacks (2017). "Current Approaches used in Epidemiologic Studies to Examine Short-Term Multipollutant Air Pollution Exposures". In: *Annals of Epidemiology* 27.2, pp. 145–153.

Dominici, F., R. D. Peng, C. D. Barr, and M. L. Bell (2010). "Protecting Human Health from Air Pollution: Shifting from a Single-Pollutant to a Multi-Pollutant Approach". In: *Epidemiology (Cambridge, Mass.)* 21.2, p. 187.

Feng, C., J. Li, W. Sun, Y. Zhang, and Q. Wang (2016). "Impact of Ambient Fine Particulate Matter (PM 2.5) Exposure on the Risk of Influenza-Like-Illness: A Time-Series Analysis in Beijing, China". In: *Environmental Health* 15.1, p. 17.

Franklin, M. and J. Schwartz (2008). "The Impact of Secondary Particles on the Association Between Ambient Ozone and Mortality". In: *Environmental Health Perspectives* 116.4, pp. 453–458.

Golchi, S., D. R. Bingham, H. Chipman, and D. A. Campbell (2015). "Monotone Emulation of Computer Experiments". In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1, pp. 370–392.

Huang, G., P. Brown, and H. H. Shin (2020). "Daily Mortality/Morbidity and Air Quality: Using Multivariate Time Series with Seasonally Varying Covariances". In: *Submitted*.

Liu, C., R. Chen, F. Sera, A. M. Vicedo-Cabrera, Y. Guo, S. Tong, M. S. Coelho, P. H. Saldiva, E. Lavigne, and P. Matus (2019). "Ambient Particulate Air Pollution and Daily Mortality in 652 Cities". In: *NEJM* 381.8, pp. 705–715.

Masselot, P., F. Chebana, C. Campagna, E. Lavigne, T. B. Ouarda, and P. Gosselin (2020). "Constrained Groupwise Additive Index Models". In: *Submitted*.

Ramsay, J. O. (1988). "Monotone Regression Splines in Action". In: *Statistical Science* 3.4, pp. 425–441.

Rue, H., S. Martino, and N. Chopin (2009). "Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations". In: *JRSS: Series B* 71.2, pp. 319–392.

Simpson, D., H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye (2017). "Penalising model component complexity: A principled, practical approach to constructing priors". In: *Statistical science* 32.1, pp. 1–28.

Stieb, D. M., R. T. Burnett, M. Smith-Doiron, O. Brion, H. H. Shin, and V. Economou (2008). "A New Multipollutant, No-threshold Air Quality Health Index Based on Short-Term Associations Observed in Daily Time-Series Analyses". In: *Journal of the Air & Waste Management Association* 58.3, pp. 435–450.

Stringer, A., P. Brown, and J. Stafford (2020a). "Approximate Bayesian inference for case-crossover models". In: *Biometrics*.

– (2020b). "Fast, Scalable Approximations to Posterior Distributions in Extended Latent Gaussian Models". Submitted.

Wei, Y., Y. Wang, Q. Di, C. Choirat, Y. Wang, P. Koutrakis, A. Zanobetti, F. Dominici, and J. D. Schwartz (2019). "Short Term Exposure to Fine Particulate Matter and Hospital Admission Risks and Costs in the Medicare Population: Time Stratified, Case Crossover Study". In: *BMJ* 367.

Wu, X., R. C. Nethery, B. M. Sabath, D. Braun, and F. Dominici (2020). "Exposure to Air Pollution and COVID-19 Mortality in the United States". In: *medRxiv*.

Xia, Y. and H. Tong (2006). "Cumulative Effects of Air Pollution on Public Health". In: *Statistics in Medicine* 25.20, pp. 3548–3559.

Anticipated roles of trainees (students and post-doctoral fellows)

Kamal Rai will complete his PhD in 2021 and the intention is for him to work on this project as a postdoc. He will develop the Bayesian implementation of the GAIM models in Stan. This includes exploring determining appropriate prior distributions for the weights α, developing visualizations that communicate modeling results. Following this, he will work on either the shape-constrained models or the INLA-like algorithm, depending on the project chosen by the PhD student. He will be responsible for producing paper(s) summarizing the results of this model when run on Canadian air pollution and mortality data. To facilitate team communication and cohesion, he will spend the first year in Toronto under the supervision of Drs. Brown and Franklin, and then relocate to Halifax to work with Drs. Feng and Rainham at Dalhousie.

A PhD student will be based at INRS in Quebec City under the primary supervision of Fateh Chebana. Depending on the interests and abilities of the student, they will work on either shape-constrained models or the INLA-like algorithm (or possible both). This student will also lead the application of the methods to the Quebec data, which cannot leave the province. This student will interact regularly with all four investigators via videoconference. At least one of the Toronto or Dalhousie investigators will be a co-supervisor and a year-long visit to Dalhousie or Toronto will be encouraged.

Each year one MSc student at Dalhousie (or possibly Toronto) will be assigned a project aligned with their interests and abilities. A more mathematically-focused project could be prior distributions for Gaussian processes which are near-monotonic, an application-driven project would be to do thorough comparison of the methods developed to existing methodologies. Co-supervision by Investigators at different institutions will be pursued where possible.

Two summer NSERC undergraduate research students per year will be engaged to work on a project closely aligned with the CRT graduate student or postoc at their institution, who will provide the day-to-day supervision.

Equity, Diversity, Inclusion

An attempt has been made to assemble a diverse team. Two of the four investigators are women and two are ethnic minority. The proposed postdoc is also ethnic minority, the collaborators include a further two women and two ethnic minorities. Several of the investigators have young children and are sympathetic to any family commitments which collaborators and trainees may have. While there are important groups which are not currently reflected in the team, the investigators are committed to making all potential team members welcome and value the range of opinions and experiences which a diverse team will bring.

The Centre for Global Health Research (CGHR) has a history of hosting visiting scholars from all parts of the globe, including many from Africa. It is both because of the links CGHR has internationally and the focus of the research which attracts an ethnically and geographically diverse group. Ongoing research projects concern sex-selective abortion, gun violence and ethnicity in several countries in North and South America, and the effects of poverty and illiteracy on mortality in India.

In advertising for trainees, we will state that applicants wishing to work with problems in

global health, marginalized communities, or sex discrepancies in health outcomes would be able to work on applications in these areas. To give one example, snakebite causes over 50,000 deaths annually in India, mostly in deprived rural areas. Data from the Million Deaths Study has shown strong seasonal variation and appears to be correlated with environmental factors. Using bcGAIM it should be possible to quantify how rainfall, temperature, and agricultural schedules affect snakebite mortality.

While it is intended that the trainees will travel between the three cities (and Ottawa), it is understandable that family or other commitments may require relying on videoconference in some cases. Prospective trainees wishing to work primarily in French could be accommodated with co-supervision by Drs. Brown and Chebana. We expect that the combination of this exciting application area and the diversity of the research team will encourage participation from individuals who currently feel excluded by the mathematical sciences research community.

Anticipated organization of collaboration

The past 6 months have demonstrated both the feasibility of remote collaborations via videoconference and the drawbacks to extended periods without in-person communication. The team will have three organizational structures: strong local relationships; extended visits by trainees; and team-wide meetings.

Each research activity will be anchored by a trainee who will have a principle supervisor at their local institution and a co-supervisor at one of the other two sites. Each of these groups will meet weekly, with the remote supervisor possibly joining bi-weekly. The trainee in question will be responsible for organizing meetings, keeping track of tasks, and circulating a summary of the discussion after each meeting.

These teams will be combined into sub-projects (e.g. computational sub-project or AQHI sub-project) which will meet monthly. Meetings of the entire CRT will happen every other month, with one yearly day-long in-person meeting of the entire group (likely before or after the Statistical Society of Canada conference). The postdoc will organize these meetings. The four Investigators will form an 'executive group' and will meet separately as needed.

The group will set up a Slack channel for messaging and a Github repository for hosting code and documents under version control. To facilitate collaboration on software and manuscripts, code will be structured in R packages (even code for internal use only) and documents prepared using R Markdown.

Plans for dissemination and communication

THIS NEEDS TO BE UPDATED

The lead investigators of this proposal have a track record of publishing research results in leading statistical and epidemiological journals, and aim to publish the results of this project in high-impact journals. The results and findings of this multiple pollutant inquiry will also be shared with Health Canada and the Institut National de Santé Publique du Québec. Drs Shin and Campagna will use the methodologies developed in their ongoing research and programatic work, and facilitate the adoption of the methods more widely in

their organizations.

Schedule of events

THIS NEEDS TO BE UPDATED

Project Milestones

Task: Develop the bcGAIM

- 6-8 months: Implement prior(s) for shape-constrained inference for 1^{st}-order and 2^{nd}-order random walks.
- 8-12 months: Implement the bcGAIM in Stan and apply to the multi-pollutant problem.
- 12-16 months: Iterate development of priors until modeling results are satisfactory for the multi=pollutant model.
- 16-20 months: Write a paper summarizing these modeling results and submit for publication.
- 16-20 months: Release an R package so that these models are readily available.
- 20-24 months: Implement additional prior(s) for shape-constrained Bayesian inference.
- 24-30 months: Extend to a hierarchical model.
- 30-36 months: Write a paper summarizing the results of the hierarhical extensions for the exact and approximate inference models, and submit for publication.

Task: Approximate Inference Algorithm

- 4-6 months: Implement Laplace approximations for $\pi(\eta|\theta, \alpha, Y)$ and $\pi(\theta|Y)$ in Stan.
- 4-8 months: Implement an approximate inference algorithm for $\pi(\alpha|\theta, Y)$ in Stan.
- 8-12 months: Implement both approximations outside of Stan. Compare estimation results to those achieved in Stan.
- 12-16 months: Iterate development of approximation schemes outside of Stan.
- 16-20 months: Write a paper summarizing these results and submit for publication.
- 20-24 months: Apply both approximations to the multi-pollutant model.
- 20-24 months: Add the approximate Bayesian inference models to the R package.
- 24-30 months: Extend the approximation algorithm to hierarchical models. Compare to results obtained by Stan.
- 30-36 months: Add hierarchical approximate inference model to the R package.

Task: Multi-Pollutant Application

Note: The Stan implementation of the bcGAIM and the approximate inference algorithms will be developed against the multi-pollutant model, so the tasks for this application are

mostly listed above. We include it on its own to give a specific breakdown of the development of the mulit-pollutant model.

- 8-12 months: Explore the performance of bcGAIM across regions and mortality outcomes.
- 12-16 months: Iteratively refine the bcGAIM (including the shape-constraining priors).
- 24-30 months: Extend the multi-pollutant model to a hierarchical model (exact and approximate versions).
- 30-36 months: Write a paper summarizing these results and submit for publication.
- 30-36 months: Add hierarhical bcGAIM model to the R package.

Task: COVID-19 Application

- 12-14 months: Identify COVID-19 confounders and data sets that may be used to fit a COVID-19 bcGAIM model.
- 14-16 months: Fit the bcGAIM model to COVID-19 mortality data.
- 16-20 months: Write a paper summarizing these results and submit for publication.

Task: Collaborative Applications

- 12-36 months: Once the bcGAIM is implemented, work with collaborations on appropriate epidemological studies.

Dissemination and Publication Activities

THIS NEEDS TO BE UPDATED

Year 1

- Milestone: Implement the bcGAIM with shape-constrained priors that are applicable for the multi-pollutant problem.
- Milestone: Implement the INLA-like approximation to the target density of this model in Stan.

Year 2

- Milestone:
- Milestone:
- Submit paper: A multi-pollutant air quality index.
- Submit Paper: Approximate Bayesian inference for the bcGAIM model
- Dissemination: Discuss shape-constrained Bayesian inference at 1-2 conferences.
- Dissemination: Discuss approximate Bayesian inference at 1-2 conferences.

Year 3

- Milestone:
- Submit Paper: Shape-constrained Bayesian inference with interpretable priors.
- Submit paper: The effects of multiple pollutant mixtures on COVID-19 mortality.
- Submit paper: A hierarchical extension to Approximate Bayesian inference
- Dissemintation: Discuss the multi-pollutant air quality index at 1-2 conferences.
- Dissemintation: Discuss hierarchical extensions to the multi-pollutant air quality index at 1-2 conferences.
- Dissemintation: Discuss the COVID-19 inquiry into air pollution exposure at 1-2 conferences.

Major Collaborative Activities

THIS NEEDS TO BE UPDATED

The different components of the bcGAIM project are naturally related. The bcGAIM is being developed in Stan in the first year, as is the first version of the approximate inference algorithm. Therefore, the bcGAIM model should be written to faciliate incorporating these approximations, and they should be developed knowing they will be implemented in Stan. In the second year, the two Stan models and the approximate inference algorithm will be extended to a hierarchical formulation. Although the hierarchical structure is at the city-level, nearby cities differ in their distance from each other. Ideally, a hierarchical model should account for how the composition of a mixture of pollutants varies by distance. The numerical difficulties and more complicated hierarchical structure should encourage strong collaboration at this stage of the project. The third year is devoted to applications – applying the fully developed bcGAIM model to the multi-pollutant problem, COVID-19 data, and other epiodemiological applications that arise during the course of the project – as well as writing papers and producing a useful R package. There is again natural collaboration between those writing and maintaining the R package.

Three-year budget

The table below lists income and expenses for each of the three years of the project. This budget assumes that CANSSI funding over the 3 year period is $200,000.

“‘{r CANSSI-Budget-Yearly-Breakdown, echo = FALSE} library(magrittr) Expenses <- t(data.frame(“\makecell[l]{Postdotor-\\al Fellow}”, “\makecell[l]{Graduate \\ Students}”, “\makecell[l]{Graduate \\ Students}”, “\makecell[l]{Undergrad \\ Students}”, “\makecell[l]{Undergrad \\ Students}”, “\makecell[l]{Research \\ Assistant}”, “\makecell[l]{INSPQ \\ Staff Time}”, “Travel”, “\textbf{Total}”))

Contributions_1_Names_Non_CANSSI <- t(data.frame(“HC”, “INRS”, “NSERC”, “USRA”, “DG”, “CGHR”, “INSPQ”, ” “,” ”))

Contributions_1_CANSSI <- t(data.frame(“\makecell[c]{\$35}”, “\makecell[c]{ \textcolor{white}{0} \$24}”, “\makecell[c]{ \textcolor{white}{0} \$24}”, “\textcolor{white}{1-1}”, “\textcolor{white}{1-1}”, “\textcolor{white}{1-2}”, “\textcolor{white}{1-3}”, “\makecell[c]{\$7}”, “\makecell[c]{\textbf{\$66}}”))

Contributions_2_CANSSI <- t(data.frame(“\makecell[c]{\$35}”, “\makecell[c]{ \textcolor{white}{0} \$24}”, “\makecell[c]{ \textcolor{white}{0} \$24}”, “\textcolor{white}{1-1}”, “\textcolor{white}{1-1}”, “\textcolor{white}{1-2}”, “\textcolor{white}{1-3}”, “\makecell[c]{\$7}”, “\makecell[c]{\textbf{\$66}}”))

Contributions_3_CANSSI <- t(data.frame(“\makecell[c]{\$35}”, “\makecell[c]{ \textcolor{white}{0} \$24}”, “\makecell[c]{ \textcolor{white}{0} \$24}”, “\textcolor{white}{1-1}”, “\textcolor{white}{1-1}”, “\textcolor{white}{1-2}”, “\textcolor{white}{1-3}”, “\makecell[c]{\$9}”, “\makecell[c]{\textbf{\$68}}”))

Contributions_1_Non_CANSSI <- t(data.frame(“HC: \space \space \space \space \space \$20”, “INRS: \space \space \$10”, “NSERC: \$7”, “USRA: \space \space \$9”, “DG:

\space \space \space \space \space \space \$3", "CGHR: \$10", "INSPQ: \$10", " ","\makecell[l]{ \textcolor{white}{0} \space \space \space \space \space \space \space \space \space \textbf{\$69}}"))

Contributions_1_Non_CANSSI[1] <- paste0("HC\textbf{", kableExtra::footnote_marker_number(1), "}", " \space \space \space \space \space \space \$20")
Contributions_1_Non_CANSSI[2] <- paste0("INRS\textbf{", kableExtra::footnote_marker_number(2),"
\space \space \space \$20")
Contributions_1_Non_CANSSI[3] <- paste0("NSERC\textbf{", kableExtra::footnote_marker_number(3
\space \$7") Contributions_1_Non_CANSSI[4] <- paste0("USRA\textbf{", kableExtra::footnote_marker_number(4), "}", " \thinspace \thinspace \space \space \$9") Contributions_1_Non_CANSSI[5] <- paste0("DG\textbf{", kableExtra::footnote_marker_number(5), "}", " \thinspace \space \space \space \space \space \space \$3") Contributions_1_Non_CANSSI[6] <- paste0("CGHR\textbf{", kableExtra::footnote_marker_number(6),"}"," \thinspace \thinspace \$10") Contributions_1_Non_CANSSI[7] <- paste0("INSPQ\textbf{", kableExtra::footnote_marker_number(7),"}"," \thinspace \thinspace \$10")

Contributions_2_Non_CANSSI <- t(data.frame("HC \space \space \space \space \space \space \space \space \$20", "INRS \space \space \space \space \space \$10", "NSERC \space \space \space \$7", "USRA \space \space \space \space \space \$9", "DG \space \space \space \space \space \space \space \space \space \$3", "CGHR \space \space \space \$10", "INSPQ \space \space \space \$10", " ","\makecell[l]{ \textcolor{white}{0} \space \space \space \space \space \space \space \space \space \textbf{\$69}}"))

Contributions_3_Non_CANSSI <- t(data.frame("HC \space \space \space \space \space \space \space \space \$20", "INRS \space \space \space \space \space \$10", "NSERC \space \space \space \$7", "USRA \space \space \space \space \space \$9", "DG \space \space \space \space \space \space \space \space \space \$3", "CGHR \space \space \space \$10", "INSPQ \space \space \space \$10", " ","\makecell[l]{ \textcolor{white}{0} \space \space \space \space \space \space \space \space \space \textbf{\$69}}"))

Budget_Table <- cbind(Expenses, Contributions_1_CANSSI, Contributions_1_Non_CANSSI, Contributions_2_CANSSI, Contributions_2_Non_CANSSI, Contributions_3_CANSSI, Contributions_3_Non_CANSSI) colnames(Budget_Table) = c(""","CANSSI","Non-CANSII","CANSSI","Non-CANSSI","CANSSI","Non-CANSSI") rownames(Budget_Table) = NULL

knitr::kable(Budget_Table, booktabs = TRUE, escape = FALSE, format = "latex", label = "Budget-Table") %>% kableExtra::row_spec(0, align = "l") %>% kableExtra::column_spec(c(1,3,5,7), border_right = TRUE) %>% kableExtra::column_spec(c(1,2), border_left = TRUE) %>% kableExtra::collapse_rows(columns = c(1,2,4,6)) %>% kableExtra::add_header_above(c(" " = 1, "Year 1" = 2, "Year 2" = 2, "Year 3" = 2)) %>% kableExtra::add_header_above(c("Expenses" = 1, "Contributions (in 1,000's)" = 6)) %>% kableExtra::footnote(number = c("HC: Health Canada", "INRS: Institut national de la recherche scientifique", "NSERC: The Natural Sciences and Engineering Research Council of Canada", "USRA: NSERC Undergraduate Student Research Award", "DG: NSERC Discovery Grant", "CHGR: Centre for Global Health Research", "INSPQ: Institut

national de santé publique du Québec (Quebec Public Health Institute)")) "'

Annual Expenses

1. Postdoctoral Fellow: The postdoctoral fellow will be funded from the CANSSI CRT grant and Health Canada. He will help organize team meetings, split time between Toronto and Halifax, and help onboard other students as they join the project.
2. Graduate Students: 1 PhD student and 1 Master's student will be involved in this project. One of the graduate students will be based at INRS, the other will be at the University of Toronto or Dalhousie University.
3. URSA Students: This project will have a number of self-contained projects suitable for undergraduate research assistants. We intend to involve 2 URSAs at $6,000/year each, whose work will directly contribute to the project's research aims.
4. Research Assistant: Hana Fu at CGHR will contribute roughly 3 days/month to the project. She will help maintain project data files and perform preliminary analysis.
5. INSPQ Staff Time: Céline Campagna at INSPQ will devote 0.5 days/week to the project.
6. Travel/Equipment: The travel expenses will cover attending conferences and travel between the three institutions by the project trainees. The equipment spending is intended to cover new computing equipment or cloud computing costs.

Contributions

1. CANSSI: The CANSSI funding is $200,000 over three years, or $66,666 per year.
2. Health Canada: Health Canada will contribute $20,000 per year via research contracts.
3. INRS: Fateh Chebana will contribute $10,000 in graduate student funding.
4. NSERC: Cindy Feng will contribute $7,000/year in graduate student funding.
5. INSPQ: The INSPQ will contribute staff time to the project, estimated at $10,000/year for 3 years.
6. CGHR: CGHR's support will be in-kind, in the form of funding the research assistant and providing research facilities.
7. Undergraduate Summer Students: Two USRAs will be applied for each year, which will pay for undergraduate summer students. NSERC also requires a contribution from the Discovery Grant of the supervisor.

References

Bobb, J. F., L. Valeri, B. Claus Henn, D. C. Christiani, R. O. Wright, M. Mazumdar, J. J. Godleski, and B. A. Coull (2015). "Bayesian Kernel Machine Regression for Estimating the Health Effects of Multi-Pollutant Mixtures". In: *Biostatistics* 16.3, pp. 493–508.

Bopp, S., A. Richarz, A. Worth, E. Berggren, and M. Whelan (2018). "Something from Nothing: Ensuring the Safety of Chemical Mixtures". In: *Ensuring the safety of chemical mixtures, Publications Office of the European Union, EUR* 29258.

Bürkner, P. and E. Charpentier (2020). "Modelling Monotonic Effects of Ordinal Predictors in Bayesian Regression Models". In: *British Journal of Mathematical and Statistical Psychology*.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software* 76.1.

Davalos, A. D., T. J. Luben, A. H. Herring, and J. D. Sacks (2017). "Current Approaches used in Epidemiologic Studies to Examine Short-Term Multipollutant Air Pollution Exposures". In: *Annals of Epidemiology* 27.2, pp. 145–153.

Dominici, F., R. D. Peng, C. D. Barr, and M. L. Bell (2010). "Protecting Human Health from Air Pollution: Shifting from a Single-Pollutant to a Multi-Pollutant Approach". In: *Epidemiology (Cambridge, Mass.)* 21.2, p. 187.

Feng, C., J. Li, W. Sun, Y. Zhang, and Q. Wang (2016). "Impact of Ambient Fine Particulate Matter (PM 2.5) Exposure on the Risk of Influenza-Like-Illness: A Time-Series Analysis in Beijing, China". In: *Environmental Health* 15.1, p. 17.

Franklin, M. and J. Schwartz (2008). "The Impact of Secondary Particles on the Association Between Ambient Ozone and Mortality". In: *Environmental Health Perspectives* 116.4, pp. 453–458.

Golchi, S., D. R. Bingham, H. Chipman, and D. A. Campbell (2015). "Monotone Emulation of Computer Experiments". In: *SIAM/ASA Journal on Uncertainty Quantification* 3.1, pp. 370–392.

Huang, G., P. Brown, and H. H. Shin (2020). "Daily Mortality/Morbidity and Air Quality: Using Multivariate Time Series with Seasonally Varying Covariances". In: *Submitted*.

Liu, C., R. Chen, F. Sera, A. M. Vicedo-Cabrera, Y. Guo, S. Tong, M. S. Coelho, P. H. Saldiva, E. Lavigne, and P. Matus (2019). "Ambient Particulate Air Pollution and Daily Mortality in 652 Cities". In: *NEJM* 381.8, pp. 705–715.

Masselot, P., F. Chebana, C. Campagna, E. Lavigne, T. B. Ouarda, and P. Gosselin (2020). "Constrained Groupwise Additive Index Models". In: *Submitted*.

Ramsay, J. O. (1988). "Monotone Regression Splines in Action". In: *Statistical Science* 3.4, pp. 425–441.

Rue, H., S. Martino, and N. Chopin (2009). "Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations". In: *JRSS: Series B* 71.2, pp. 319–392.

Simpson, D., H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye (2017). "Penalising model component complexity: A principled, practical approach to constructing priors". In: *Statistical science* 32.1, pp. 1–28.

Stieb, D. M., R. T. Burnett, M. Smith-Doiron, O. Brion, H. H. Shin, and V. Economou (2008). "A New Multipollutant, No-threshold Air Quality Health Index Based on Short-Term Associations Observed in Daily Time-Series Analyses". In: *Journal of the Air & Waste Management Association* 58.3, pp. 435–450.

Stringer, A., P. Brown, and J. Stafford (2020a). "Approximate Bayesian inference for case-crossover models". In: *Biometrics*.

Stringer, A., P. Brown, and J. Stafford (2020b). "Fast, Scalable Approximations to Posterior Distributions in Extended Latent Gaussian Models". Submitted.

Wei, Y., Y. Wang, Q. Di, C. Choirat, Y. Wang, P. Koutrakis, A. Zanobetti, F. Dominici, and J. D. Schwartz (2019). "Short Term Exposure to Fine Particulate Matter and Hospital Admission Risks and Costs in the Medicare Population: Time Stratified, Case Crossover Study". In: *BMJ* 367.

Wu, X., R. C. Nethery, B. M. Sabath, D. Braun, and F. Dominici (2020). "Exposure to Air Pollution and COVID-19 Mortality in the United States". In: *medRxiv*.

Xia, Y. and H. Tong (2006). "Cumulative Effects of Air Pollution on Public Health". In: *Statistics in Medicine* 25.20, pp. 3548–3559.