Using mixed model approaches can leverage database information to improve the estimation of size-adjusted contaminant concentrations in fish populations

\* Emily Smenderovac (corresponding author), Great Lakes Forestry Centre, Natural Resources Canada, [emily.smenderovac@nrcan-rncan.gc.ca](mailto:emily.smenderovac@nrcan-rncan.gc.ca)

Brian W. Kielstra, Ecometrix, Guelph, Ontario, [bkielstra@ecometrix.ca](mailto:bkielstra@ecometrix.ca)

Calvin Kluke, Living with Lakes Centre, Laurentian University, [calvin.kluke1@gmail.com](mailto:calvin.kluke1@gmail.com)

Tom Johnston, Living with Lakes Centre, Laurentian University, Ontario Ministry of Natural Resources and Forestry, [tjohnston@laurentian.ca](mailto:tjohnston@laurentian.ca)

Satyendra P. Bhavsar, Ontario Ministry of the Environment and Climate Change, [satyendra.bhavsar@ontario.ca](mailto:satyendra.bhavsar@ontario.ca)

Rob Mackereth, Centre of Excellence for Sustainable Mining and Exploration, Ontario Ministry of Natural Resources and Forestry, [rob.mackereth@ontario.ca](mailto:rob.mackereth@ontario.ca)

Stephanie Melles, Department of Chemistry and Biology, Toronto Metropolitan University, [stephanie.melles@torontomu.ca](mailto:stephanie.melles@torontomu.ca)

Gretchen L. Lescord, Forests Fisheries and Geomatic Sciences, University of Florida, [lescord.g@ufl.edu](mailto:lescord.g@ufl.edu)

Erik J.S. Emilson, Great Lakes Forestry Centre, Natural Resources Canada, [erik.emilson@nrcan-rncan.gc.ca](mailto:erik.emilson@nrcan-rncan.gc.ca)

2024-04-30

# 1 Abstract

The concentration of bioaccumulative contaminants in fish increase with their size and age. Thus, research and monitoring on these contaminants across ecosystems is made more confounded by size effects. To account for these patterns, size-standardization of contaminant concentrations within fish populations is a common practice. Standardized concentrations are often estimated using within-population regression models (referred to as sampling event regressions, or SERs, herein). This approach requires higher sample sizes than mixed effect models, which are not as commonly used for this purpose. Herein we compare SERs to three mixed effects modeling approaches; restricted maximum likelihood (REML), Bayesian inference via Markov Chain Monte Carlo (MCMC) and, approximate Bayesian inference with nested Laplace approximation (INLA). We did this for two contaminants: mercury (Hg), a contaminant known to bioaccumulate, and arsenic (As), where the bioaccumulative potential is less understood. The mixed model approaches generated accurate size-standardized concentrations for small populations (e.g., <5 fish) and/or populations which lacked the range of sizes required for SER estimates. INLA was determined to be the best method in most cases, because it was computationally less intensive than other approaches and showed consistent performance in a range of sample-size limitation scenarios. Herein, we have provided example code for prediction using the R-INLA package to enable it’s use and application in fisheries contaminant monitoring and research

# 2 Key Words

Mercury, Arsenic, Fish, Mixed effects models, contaminant modelling

# 3 Introduction

Older and larger fish tend to have higher concentrations of bioaccumulative contaminants that accumulate faster than can be excreted (e.g., mercury, polychlorinated biphenyls). Contaminant monitoring programs often standardize concentrations of these contaminants to common ages, lengths, or weights across fish populations when making comparisons or dietary recommendations, such as in the Guide for Eating Ontario Fish1. Contaminant standardization is also used to control for fish size differences among study sites in research analyses (e.g., Investigations into drivers of contaminant concentrations due to land use, forest change, and natural variation2–5).

A common approach for standardization of fish contaminant concentrations in both research and monitoring programs is to generate a linear equation between concentration and a body size metric, and use this to estimate concentration at a chosen representative size (e.g., 500 g or 1000 g) for each population1,4,6–8. This approach (herein referred to as Sampling Event Regressions (SER)) is applied to a broad range of monitored contaminants that can vary in bioaccumulative potential based on chemical characteristics, location, and fish species1. The limitation of the SER approach is that it requires multiple fish (e.g., 5+ fish/population9) across a suitable range of body sizes, which can be cost- or time-restrictive for monitoring or research programs. Smaller sample sizes can lead to poor inferences10 and/or exclusion of underrepresented waterbodies or species when using the SER approach11.

A potential alternative to SER, without the sample size limitations, are mixed effects modelling approaches. These approaches offer the advantage of ‘borrowing’ from all observations across sites for estimating overall “average” relationships while using deviation in absolute values (random intercepts) or relationships (random slopes) among sites (waterbodies) to generate predictions and confidence intervals for a given waterbody and a given fish size12,13. This approach can potentially increase accuracy and allow estimation even when sample sizes are low, or when there is an inadequate range of body sizes for estimation with SER. A likely barrier to uptake of these approaches by practitioners, however, is the intimidating array of possible approaches with varied complexity, accuracy, and computational requirements. The most accessible implementation is the restricted maximum likelihood approach (REML), which is relatively easy to implement, but is limited in that it does not account for uncertainty in the random effects14, which could reduce model fit and increase uncertainty in estimated concentrations. Relatively user-friendly, accessible statistical packages have been introduced that facilitate the inclusion of confidence intervals around estimates. This included bootstrapped likelihoods generated from REML, or estimation of probability distributions either with a Bayesian Markhov Chain Monte Carlo (MCMC) approach or the “approximate Bayesian” Integrated Laplace approximation (INLA) approach. Briefly, these latter two approaches differ in how they estimate probability distributions with MCMC using randomized resampling of both samples and model parameters15, and INLA using the (faster) Laplace approximations of probability distributions for each parameter16,17.

Herein we compare four statistical approaches to determine the most accurate way to perform contaminant concentration standardization by fish body size, particularly under data limited scenarios. More specifically, we compared the commonly applied SER to four mixed effects model approaches that differ in how they estimate posteriors: 1) a REML approach, 2) a REML approach with bootstrapping (REML\_boot), 3) a common Bayesian inference approach (MCMC), and 4) an approximate Bayesian inference (INLA). We evaluated these approaches using two inorganic contaminants of concern in fish tissue that differ in bioaccumulative potential: mercury (Hg) and arsenic (As). We hypothesized that mixed model approaches would provide comparable or more accurate predictions of contaminant concentrations relative to SERs, but allow estimation in data-scarce populations (i.e., when sample sizes are low and/or when data is limited across size ranges). Amongst mixed model approaches, we expected Bayesian approaches (i.e., MCMC and INLA) to have higher predictive accuracy than REML approaches because of the inclusion of uncertainty in random effects. We expected the REML\_boot model to have improved prediction compared to the REML models for the same reason, though with less improvement than Bayesian approaches as the bootstrapping is a more ‘brute force’ approach to error estimation.

# 4 Materials and Methods

## 4.1 Data

### 4.1.1 Fish contaminant data

We obtained a dataset of total Hg and As concentrations ([Hg] and [As] respectively) in fish muscle data from the Ontario Ministry of Environment, Conservation, and Parks (OMECP)18 and from data collected in Northern Ontario for Arsenic studies19. Although individual fish sampling protocols varied, muscle tissue sent to OMECP is analyzed using standard methods as part of the Fish Contaminants Monitoring Program (FCMP)18. More specifically, total [Hg] were measured using cold vapor-flameless atomic absorption spectroscopy (CV-FAAS) following protocol OMECP-HGBIO-WS057 and total [As] was measured using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) following method OMECP-BIOTA-E3461.

We included [Hg] because it is a highly bioaccumulative contaminant of global concern that is routinely monitored in fish and used to inform consumption guidelines20. Size adjustments of [Hg] concentrations in research of variation across environmental gradients are also common3,7,21–23. We also include [As] as a representative contaminant that has also been size-adjusted in research3,19, but is less bioaccumulative in fish, showing mixed or weaker relationships with metrics of body size3,19,24. We further limited data to only inland lakes (i.e., excluding the Great Lakes) and to three fish species: *Salvelinus namaycush namaycush* (common lake trout, hereafter lake trout), *Esox lucius* (northern pike), and *Sander vitreus* (walleye). These species were selected because while they are all predators they represent distinct niches within lakes. These species are also commonly consumed by anglers, so accurate estimation of contaminant concentrations is important for mitigating human health risk. Our final dataset included [Hg] in 37923 fish and [As] in 1001 fish. The datasets were heavily biased to smaller fish, but had a normal distribution of values, with fish between 400 and 1100 being well represented in the data (Figure S4). Generally, most fish in the dataset had contaminant concentrations < 1.3 ug/gram~wet weight~ [Hg] and < 0.4 ug/g~wet weight~ [As] (Figure S1).

### 4.1.2 Test and training datasets

To facilitate the comparison of standardization approaches, we randomly sampled the fish contaminant datasets to create test and training datasets. We focused on the ability to predict concentrations for two common weight standardization targets of 500g and 1000g, as these were round numbers well represented in the dataset, and commonly used in research studies3,6,7. As such, we randomly sampled and assigned 50% of contaminant data from fish weighing 500g ± 100g and 1000g ± 100g to the test dataset, with the remaining 50% of data in those ranges and all data outside of those weight ranges assigned as a training dataset (Table S1).

## 4.2 Standardization Approaches

### 4.2.1 Sampling event regressions (SER)

For each population of fish representing a unique contaminant-species-waterbody-year combination (i.e., a sampling event), we developed log-log (log-contaminant (in ug/g) by log-weight (in g)) regression models. SERS were limited to cases where there were a minimum of 5 sampled individuals, as used in Kluke et al, 202319, and also where the range of fish sizes allowed interpolative prediction (i.e., prediction of test set concentrations for weights inside the range of the training set). Our data filtering may be more lenient when compared to methods used in research, due in part to limited data and efficiency. Each model used the linear equation outlined in Equation (1).

where *i* is an individual fish for each species-waterbody-year combination event *k*, β0k is the intercept, β1k is the slope, and ϵik is the residual error; these models were run on training data until all combinations were exhausted.

### 4.2.2 Mixed effects models (REML, REML\_boot, MCMC and INLA)

For each species and contaminant combination, we developed log-log (log-contaminant (μg/g wet weight) by log-weight (g)) linear mixed effects models including all waterbodies and years. We allowed for random variation in the slope and intercept per waterbody, and random variation in the intercept by sampling event (i.e., each unique year and waterbody combination). The general model structure was (2).

where *i* is an individual fish in waterbody *j* during sampling event *k*, β0 is the fixed intercept, β1 is the fixed slope, u0j is the random intercept for waterbody *j*, u1j is the random slope for waterbody *j*, u0k is the random intercept for sampling event *k*, and ϵijk is residual error. This identical model structure was fit in all three mixed effects model approaches described below.

The REML models were fit using the *lme4* v1.1.35.3 package *lmer* function with default settings25 and the REML\_boot analysis was performed using *lme4*’s *bootMer* function for parametric bootstrapping using 2000 simulations and with the use.u setting set to TRUE to simulate spherical random effects using 4 CPUs. The MCMC models were fit in STAN v2.32.2 through *rstanarm* v2.32.1. All models were run with 4 chains (and 4 CPUs)15 and the priors were the coefficients of an equivalent glmer model. For [Hg], we used the default settings of 2000 iterations per chain and the adapt\_delta setting of 0.8. For [As], however, the lower sample sizes required 9000 iterations and an adapt delta setting of 0.99 for model convergence (at the cost of computation time). The INLA models were run in the *R-INLA* package with the “iid2d” model structure to account for covariance of the random effects due to waterbody26. The median of predicted test values was used as a comparison to measured values, and also calculate root mean squared error (RMSE) for all models.

## 4.3 Population medians (no standardization)

We calculated the population median of [As] values for each of the three fish species and both contaminants, individually for each waterbody-year combination (i.e., a sampling event), as it is possible that there is no bioaccumulation of this contaminant in some species. We included this comparison as a predictive approach, as good prediction with median values rather than with standardized concentrations implies that there is a lack of bioaccumulation.

## 4.4 Comparing model fit and predictive accuracy

To assess the predictive accuracy of each approach, we compared predictions of test set values to the recorded measured values using Pearson correlation with the cor.test function in R. We also assessed the fit of each approach by calculating the root mean squared error (RMSE) of the training set for each sampling event (distinct year-waterbody combination) in the dataset. Accuracies based on test data reflect the generalized predictive capability of a model, while fit reflects the accuracy of the model in representing the trends present in the training data. Additionally, to provide context about the accessibility of the different approaches, we collected and compared the required computational time required to run each model to the RMSE of the overall models.

We ran an additional analysis to assess how model performance changes with different fish and lake number sampling scenarios. We trained REML and INLA models on randomly sampled training sets with a range of lake numbers and fish numbers, in replicates of 10, to evaluate model predictive performance. We used the less computationally intensive INLA approach to represent both Bayesian approaches, as INLA and MCMC were highly similar. Each randomized sample pulled a set number of lakes with a minimum number of fish, This was run for a range of lake numbers up to the maximum number of lakes in the dataset (2, 3, 5, 7, 9, 10, 15, 20, 30, 40, 50, 75, 100, 150, 200, 250, 300, 500, 750, 1000) and a range of fish sizes up to the maximum number of fish (3-20, 25, 30, 35, 40, 45, 50). We assessed these models based whether they had a predictable behaviour (increased or maintained performance with increases in number of lakes and fish) based on performance criteria of all simulations in the set having at least 0.75 r and the average r being equal to or exceeding 0.8.

Additionally, we collected and compare the required computational time for each approach. Computational times as presented are in the context of analyses performed in R version 4.3.3 (2024-02-29 ucrt)27 on a Dell Latitude 5510 PC running Windows 10 Enterprise with a 1.70GHz, 2208 MHz 4 Core(s), 8 logical processor, 16 GB of physical memory and 34.6 GB of virtual memory. All data transformations, summarizations and graphing were performed with the *tidyverse* v2.0.0, *ggplot2* v3.5.0 and *ggpubr* v0.6.0 packages28–30.

# 5 Results and Discussion

Of all the approaches tested in this paper, REML and INLA were the most practical for producing accurate size-standardized estimates of contaminant concentration for distinct waterbodies. Both techniques allowed for predictions in waterbodies where SER could not be performed. The predictions were highly comparable to the measured values and had favourable computational speed when compared to other approaches. REML may be more appropriate for many situations, as the implementation was less conceptually complex and more easily programmed at the time of writing this paper. However, INLA performed effectively on all datasets, and had more predictable performance when run on a variety of simulated sampling scenarios.

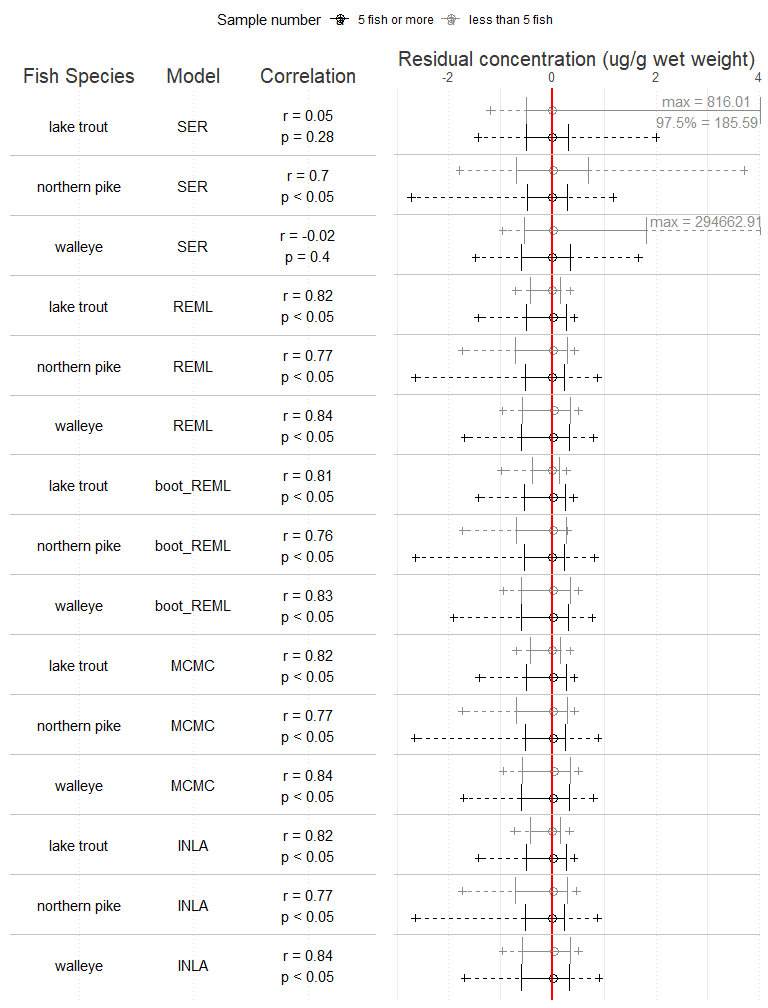


Figure 1: Table of [Hg] model prediction results for tissue of 400-600g and 900-1100g fish compared to measured values. The table includes a column of correlation statistics (r and p-value) for each model assessed in the study. To the right of the plot, a dot plot shows the distribution of model residuals for events with less than five fish (grey), or 5 or more fish (black). A central hollow circle designates the median, with a solid error bar designating the 2.5, and 95th percentiles of the distribution. A dashed line ending in a + is used to show the range of remaining residuals. A vertical red line is used to highlight where the distributions overlap with zero. Particularly large outliers or percentiles are indicated in text on the left side of the plot.

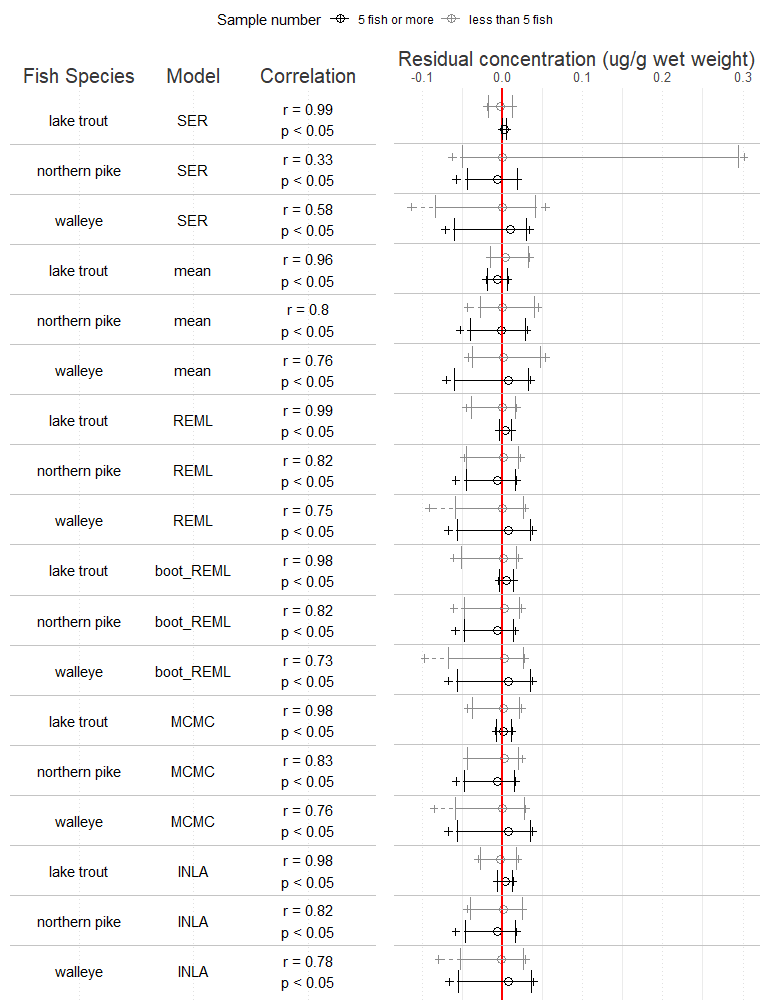


Figure 2: Table of [As] model prediction results for tissue of 400-600g and 900-1100g fish compared to measured values. The table includes a column of correlation statistics (r and p-value) for each model assessed in the study. To the right of the plot, a dot plot shows the distribution of model residuals for events with less than five fish (grey), or 5 or more fish (black). A central hollow circle designates the median, with a solid error bar designating the 2.5, and 95th percentiles of the distribution. A dashed line ending in a + is used to show the range of remaining residuals. A vertical red line is used to highlight where the distributions overlap with zero.

We found that mixed effects approaches performed better than SER by enabling prediction of values for low sample sizes with increased or comparable accuracy to SER (Figure 1, 2, Figure S2). The REML, MCMC and INLA values were generally more strongly correlated to actual concentrations (r ranging from 0.73 - 0.99) than predictions from SER models (r ranging from 0.48 – 0.84) (Figure 1, 2, Figure S2). Also, the residuals for the different mixed modeling approaches were generally similar or smaller than those of the SER approach, as observed the distributions of residuals shown in Figures 1 and 2. Interestingly, attempts to improve the REML approach with bootstrapping did not increase accuracy of predictions, even resulting in decreased correlation strengths in some cases (specifically for the [Hg] northern pike and lake trout models, and the [As] walleye model) (Figure 1 and 2). This is likely because the bootstrapping process overfit the models, as demonstrated by decreased RMSE (Figure 3), which reduced the ability of the models to produce generalized predictions. All of the mixed modeling methods were able to achieve the same level of accuracy in lakes with <5 fish as in larger sampling events, effectively increasing the size of the dataset that could produce predictions. It is likely that the increase in correlation of the mixed modelling approaches may just be a result of the increased sample size these approaches offer. Though the mixed model approach was only required for improving prediction when there was known bioaccumulation (i.e., for [Hg]), the accuracy of mixed model approaches were always comparable or better than SER or basic population means. Mixed model approaches improved prediction for datasets with less than 6000 training fish, which would be applicable to many studies. Nevertheless, the mixed model approaches are a suitable solution for contaminant prediction in a variety of dataset sizes and contaminants of variable bioaccumulative potential.

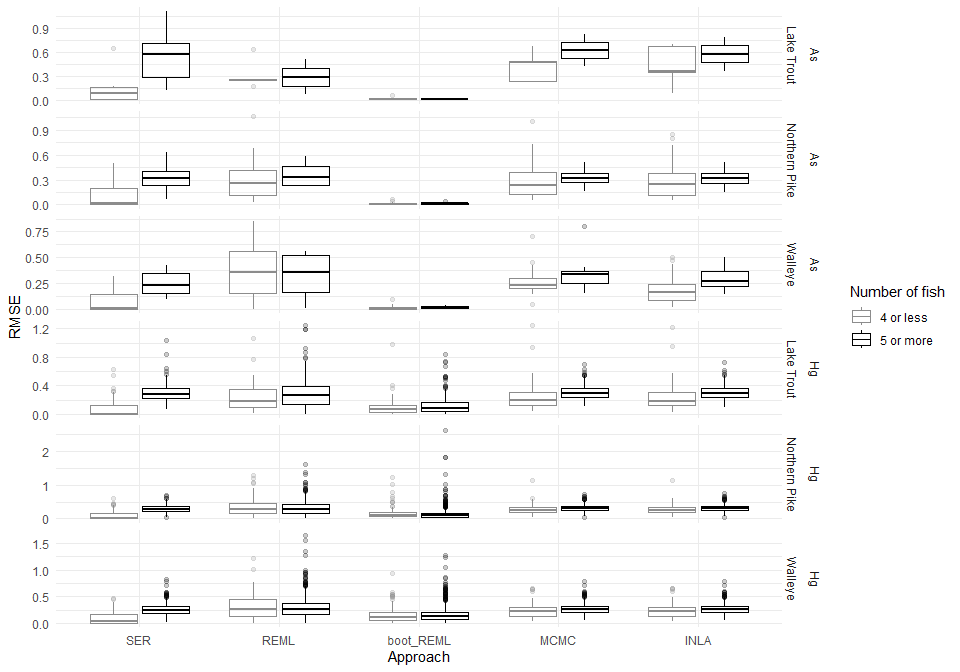


Figure 3: Root mean squared error (RMSE) from sampling events trained with different numbers of fish. RMSE of training set sampling events are compared to the number of fish used from the sampling event that were included in the training dataset. For events with less than 5 fish, and events with 5 or more fish, a boxplot displays the 25th and 75th percentiles and the median RMSE, with the whiskers extending out to 1.5 x the interquartile range. The figure shows data for [As] and [Hg] Sample event regressions (SER), restricted maximum likelihood (REML), restricted maximum likelihood with bootstrapping (boot\_REML), Markhov Chain Monte Carlo (MCMC) and Integrated Laplace Approximation (INLA) models.

Differences in the fits of values to the different sampling events illustrated how the well the different approaches could explain the data in that training set. INLA and MCMC models were more tightly fitted to the training data for the distinct sampling events, shown by lower RMSE, with fits comparable to those of SER (Figure 3). Conversely, REML approaches generally had distributions with a wider range of RMSE, including more large RMSE points for various sample-sizes. These high influential values indicated that the patterns in some sampling events were more poorly represented than others in the global model. Boot\_REML models often had the highest fits to the training data, with lower RMSE than the other mixed model approaches or SER, but this did not correspond to the best predictive accuracy. The low sample number events typically had lower RMSEs, presumably because fewer values were fit by the mixed model. All approaches had RMSE that generally increased and stabilized at about 6 fish sampling events but had fairly comparable RMSE at event sizes below 6, with the exception of some events with substantial influential points (Figure S3). These results suggest that a minimum sample size of 6 fish/population should be an effective lake sample size for future contributions to the existing databases to ensure good representation of each lake. MCMC and INLA generally appeared to have fewer large influential RMSE values, and a more even distribution of RMSE for the different sampling events. The improved performance of the MCMC and INLA approaches do not appear to have these limitations, they are more adept at tuning the lake-level slope and intercept parameters, and thus improved precision of the model fit8,31–33. Understanding the model fit of predictive approaches is not always necessary to ensure predictive accuracy, as higher fits can be higher due to exclusion of influential values or lack of balance in a dataset. However, many researchers are interested in the relationships involved with these predictions. Size-concentration relationships can provide insight into bioaccumulation differences between waterbodies. The improved fit of the INLA and MCMC approaches showed that these approaches were better at explaining the patterns in the training data than REML, and these increased fits had little or no cost to predictive accuracy (i.e., the models are better at predicting the general population trends). Thus, if a researcher is interested in the mechanisms of prediction, the Bayesian approaches would be suitable options.

## 5.1 Relative computational requirements

Table 1: RMSE for the global model for each modeling approach compared to the computation time required to perform it. The time in seconds, is presented for each modelling approach for each species and contaminant combination. Computation time ranged from less than one second to 47248 seconds (13.1 hours) across approaches and species.

| Contaminant | Species | SER | REML | boot\_REML | MCMC | INLA |
| --- | --- | --- | --- | --- | --- | --- |
| As | lake trout | 0.07 | 0.20 | 50.39 | 322.53 | 7.13 |
| As | northern pike | 0.38 | 0.39 | 57.10 | 3075.72 | 2.46 |
| As | walleye | 0.38 | 0.33 | 43.90 | 2430.68 | 2.75 |
| Hg | lake trout | 3.37 | 0.94 | 326.28 | 10722.27 | 5.40 |
| Hg | northern pike | 7.69 | 1.59 | 872.42 | 27989.16 | 12.64 |
| Hg | walleye | 8.07 | 1.78 | 750.76 | 47248.49 | 14.51 |

While our results show the benefit of using Bayesian modeling approaches when size-standardizing contaminant concentrations in fish, they are more complex and thus, can have higher computational requirements. Overall, INLA models were much less computationally intensive compared to the MCMC and REML\_boot implementations. While REML was the quickest of all modeling approaches, INLA was a close second, with times orders of magnitude lower than REML\_boot or MCMC (Table 1). As the overall accuracy of the INLA approach was improved compared to REML approach, and it had improved speed compared to the MCMC approach we tested in this study, we suggest that INLA is the most viable alternative to standardizing values via the SER approach, that may allow for incorporation of more data into analyses.

In our simulated samplings, we found that INLA and REML both performed well on large numbers of lakes, and were able to generalize the results across lakes with low sample sizes. Generally, in these balanced random samples, REML (Figure S5, S6, S7, S11, S12, S13) performed better than INLA models (Figure S8, S9, S10, S14, S15, S16). This contrasted with the results from the assessment of the full dataset, where REML and INLA models performed comparably in terms of correlation scores. These results also suggest that while these models can decrease the number of fish required to get a good estimate for a single lake, that they require data from a large amount of lakes. Northern pike and lake trout mercury prediction was possible in simulations where 4 fish were used, but these predictions required more than 100 lakes (Figure S11, S12, S14, S15). For example, lake trout prediction of mercury required models with over 200 fish to predict using four fish per lake. Arsenic results were more varied, with few combinations passing our quality thresholds for acceptable performance stability and correlation strength. However, none of the arsenic datasets had very large number of lakes, which limited this line of investigation. It should be noted, that while REML models generally had more models that passed our quality thresholds, INLA models had a pattern within their correlation scores that was more consistent with the common understanding of sampling methods, i.e., that increasing sampling will increase predictive power. This pattern was particularly apparent in the Walleye mercury dataset, where the highest average correlation scores for REML models were in lakes with 10-12 fish (Figure S13), while INLA models generally had an increase in average correlation with increases in number of fish or number of lakes (Figure S16). As the REML models had slightly lower performance in models created from the whole dataset (Figure 1), this suggests that the REML models are more sensitive to sampling imbalance than INLA models. While REML produced comparable numbers of models that passed our quality thresholds, the more predictable performance of INLA suggests this approach would be effective when used with any large dataset. INLA models had the highest performance (in terms of r and RMSE) when trained on the full dataset, and upon investigation of per-event residuals, it did not appear that the predictive performance was biased to a particular sample event (Figure S4). This result showed that INLA models were robust to imbalances in the data they were trained with. We suggest creating predictive models that leveraging the power of existing large datasets with hundreds of lakes, such as the MECP database used in this study, would be the most effective way to employ an INLA standardization approach.

Each modelling approach had different benefits and drawbacks in terms of implementation. Application of each approach varied in difficulty, going from REML<SER<MCMC<INLA. The REML approach was simple enough that any beginner R user could implement it quickly, without requiring knowledge of loops or other programming. The SER approach required more programming knowledge in order to implement it efficiently, and a significant amount of pre-screening of data to ensure that data used for linear models covered a large enough span to develop a curve. There were more documentation and reading requirements to implement the MCMC and INLA models. These requirements mean there is a higher risk of incorrect implementation of INLA models with the current INLA package implementation, since the model formula input deviates from the typical R notation. There is thorough documentation of the model notation in the INLA package, but it is nonetheless a consideration in the implementation of this approach, as it added complexity when assessing the nested random effect of waterbody and sampling event. The [As] models were generally more difficult to fit due to lower sample sizes, requiring different settings, while [Hg] models ran well on the default settings of the lmer, stan\_glmer and inla functions. Conceptually, the SER approach is the most simple to understand, given that it is essentially a group of independent linear models. The REML, MCMC and INLA are all more conceptually challenging, however, there is quite a jump in the complexity of a Bayesian model from the REML approach. Comparatively, understanding of the Bayesian (MCMC), and the approximated Bayesian (INLA) approach are again more complex, which should be considered when model tuning is required, which would likely be required for smaller datasets (<300). Relatively few studies19,34,35 have used Bayesian approaches to account for size effects, compared to more traditional linear modelling. This may be, in part, due to the lack of guidelines for the best practices during implementation. For researchers pursuing predictive accuracy alone, the REML approach may be the best option due to the ease of implementation. Researchers interested in the underlying effects driving the predictions may find the additional effort to use INLA or MCMC models worthwhile, as these models had stronger fits to the training data while maintaining predictive accuracy. Additionally, the information from these models can inform sample size adjustments: intercepts can inform if the number of samples are lacking and slopes can indicate if the range of sizes are insufficient.

Mixed model approaches show promise for increasing predictive power for estimating contaminant concentrations in fish by leveraging the predictive power from historical sampling events across time and space. We found that REML MCMC, and INLA models were effective tools to predict contaminant concentrations with models of contaminant-size relationships of three fish species for two contaminants with different bioaccumulative potentials. Though the improvement on performance was much more pronounced on the smaller datasets, these techniques have more consistent performance across sample sizes, and are a good option when the bioaccumulative potential of a contaminant is unknown. The REML was the most easily implemented However, this approach was sensitive to sample size differences. Based on our simulation experiment, we found that the REML models performed better than INLA for some models on balanced data. However, REML also had poorer model fit to the training data, meaning it is less suitable for explaining the effects driving the size-contaminant relationships. INLA can be performed quickly, without high performance computing requirements. INLA and REML could both achieve comparable accuracy models with low numbers of fish in the large datasets used in this study, but had unreliable performance on simulated sampling using smaller numbers of lakes. Additionally, improvement of model performance for REML models did not follow an expected pattern, suggesting these models were more sensitive to underlying patterns in the training data distributions than INLA. In summary, we suggest INLA is a suitable option for database-supplemented size-adjustment of contaminant concentrations in fish with the potential to increase sample sizes. Use of this tool can enable greater power in investigations of important environmental drivers of contaminants in fish. We have provided code and data to assist others to use the model structure in this paper to produce size-standardized values for a greater number of study lakes.

# 6 Acknowledgements

# 7 Data Accessibility

The processed data, scripts used to create this manuscript, and the INLA example code (run\_INLA\_size-standardization.R) are available for download at the public repository managed by the WETlab at the Great Lakes Forestry Centre: <https://github.com/GLFC-WET/HGAS_master>

# 

# 8 Supplemental Tables

Table 2: Summary of number fish for each contaminant and species in the dataset used in this analysis. Values are listed as ‘train / test’ for each dataset.

| CONTAMINANT | SPECIES\_NAME | SER | REML | boot\_REML | MCMC | INLA |
| --- | --- | --- | --- | --- | --- | --- |
| As | Lake Trout | 65(10)/10(7) | 120(27)/10(7) | 120(27)/10(7) | 120(27)/10(7) | 120(27)/10(7) |
| As | Northern Pike | 156(19)/47(33) | 465(104)/50(35) | 465(104)/50(35) | 465(104)/50(35) | 465(104)/50(35) |
| As | Walleye | 152(21)/40(29) | 416(98)/43(30) | 416(98)/43(30) | 416(98)/43(30) | 416(98)/43(30) |
| Hg | Lake Trout | 4234(396)/558(312) | 6122(638)/559(313) | 6122(638)/559(313) | 6122(638)/559(313) | 6122(638)/559(313) |
| Hg | Northern Pike | 10095(1017)/1620(854) | 13870(1474)/1625(859) | 13870(1474)/1625(859) | 13870(1474)/1625(859) | 13870(1474)/1625(859) |
| Hg | Walleye | 14246(1172)/2551(1043) | 17931(1445)/2555(1047) | 17931(1445)/2555(1047) | 17931(1445)/2555(1047) | 17931(1445)/2555(1047) |

# 

# 9 Supplemental Figures

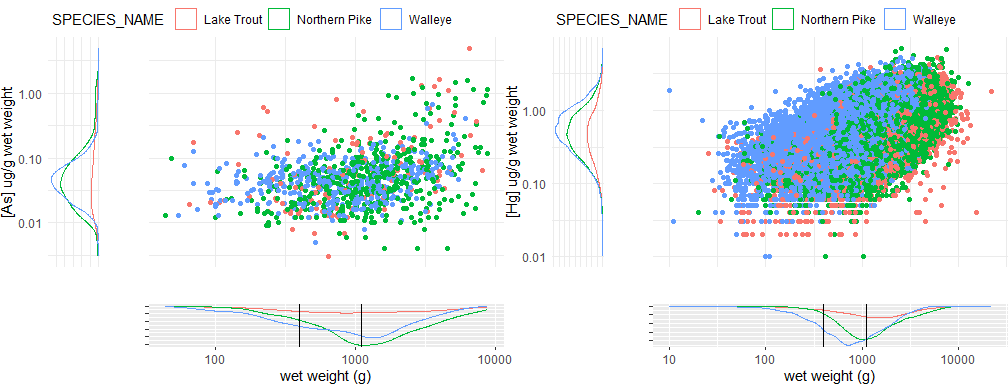


Figure 4: Number of fish at different body weights and contaminant concentrations for [As] and [Hg]. Vertical black lines show the weights that are commonly used for standardization.

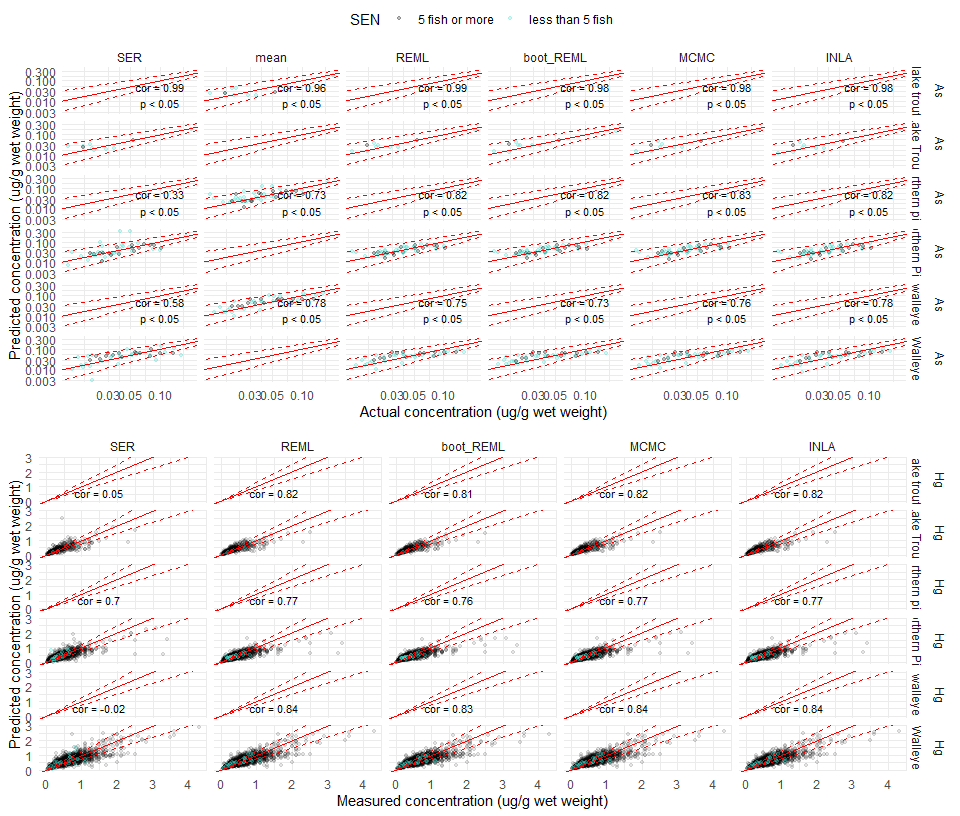


Figure 5: Predicted concentration of [As] and [Hg] in tissue of 400-600g and 900-1100g fish compared to measured values. Points display each individual fish prediction (with blue points representing predictions at sampling events with less than 5 fish, and black points representing predictions from sampling events with at least 5 fish) and a solid red line shows a 1-1 relationship. The dashed red lines provide an interval for predicted values that are within 25% of the measured value. Correlation statistics are included where there were enough values to compute them. Axes are shown on a log scale, as values were modelled from log-transformed values.

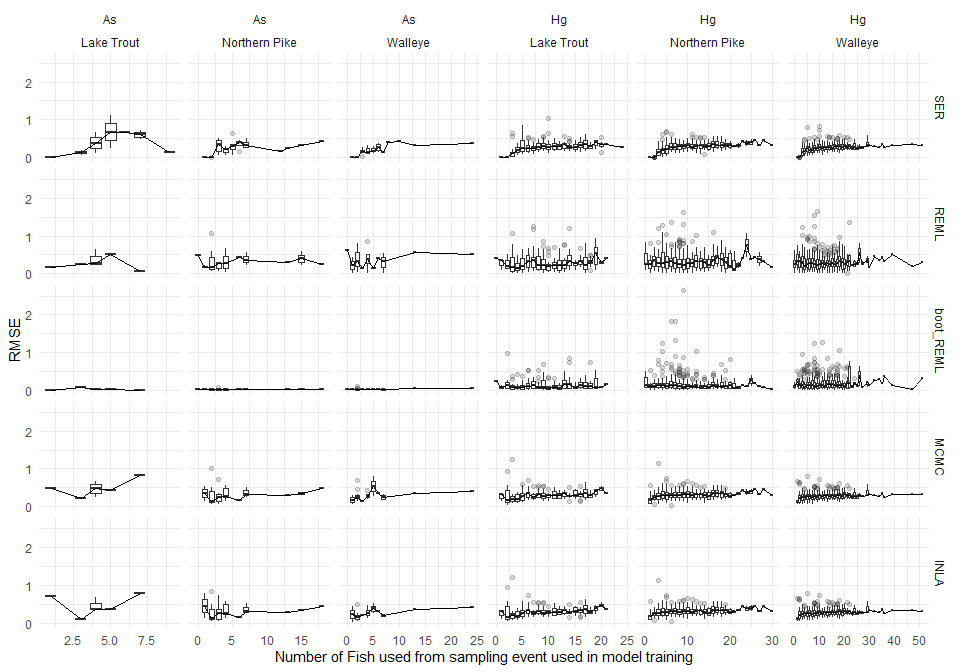


Figure 6: Root mean squared error (RMSE) of models from sampling events trained with different numbers of fish. RMSE of training set sampling events are compared to the number of fish used from the sampling event that were included in the training dataset. For each N, a boxplot displays the 25th and 75th percentiles and the median RMSE, with the whiskers extending out to 1.5 x the interquartile range. A line is plotted across the median values to show the overall trend in the values.

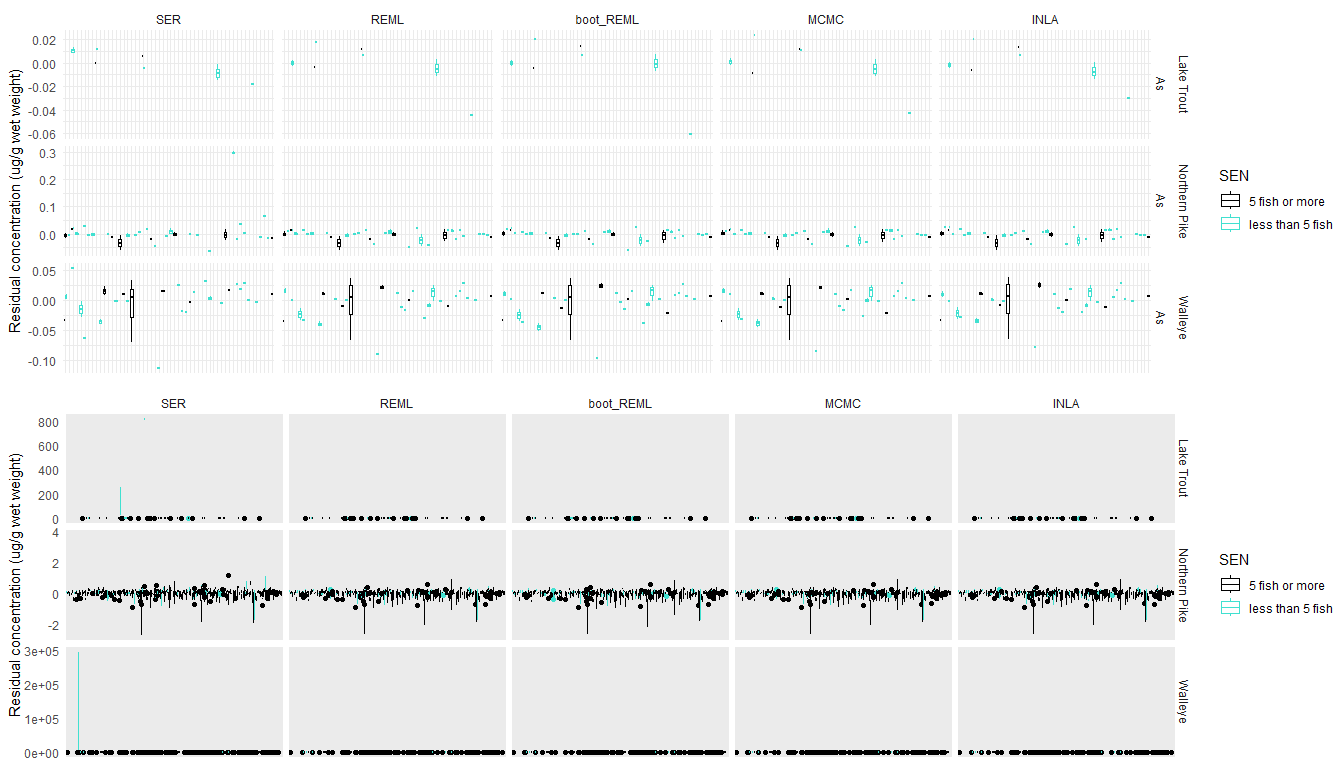


Figure 7: Residuals of predicted fish values for each sample event.

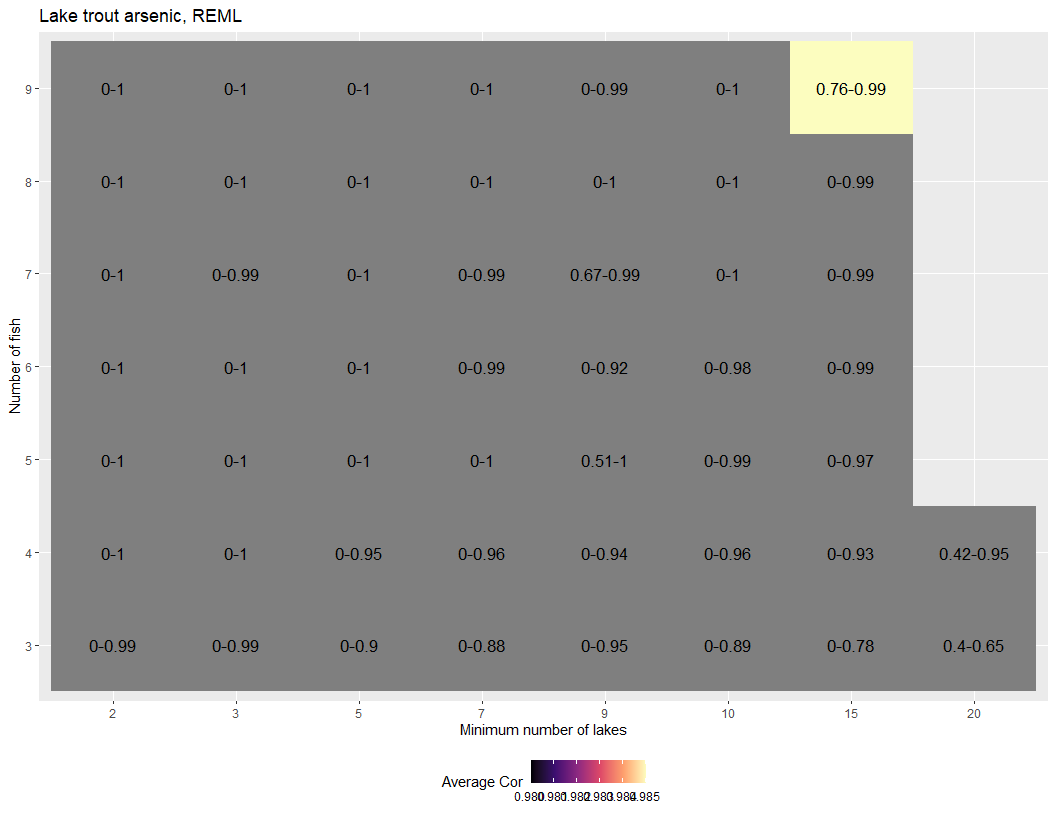


Figure 8: Range of Pearson correlation scores for simulated samplings of lake trout arsenic REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

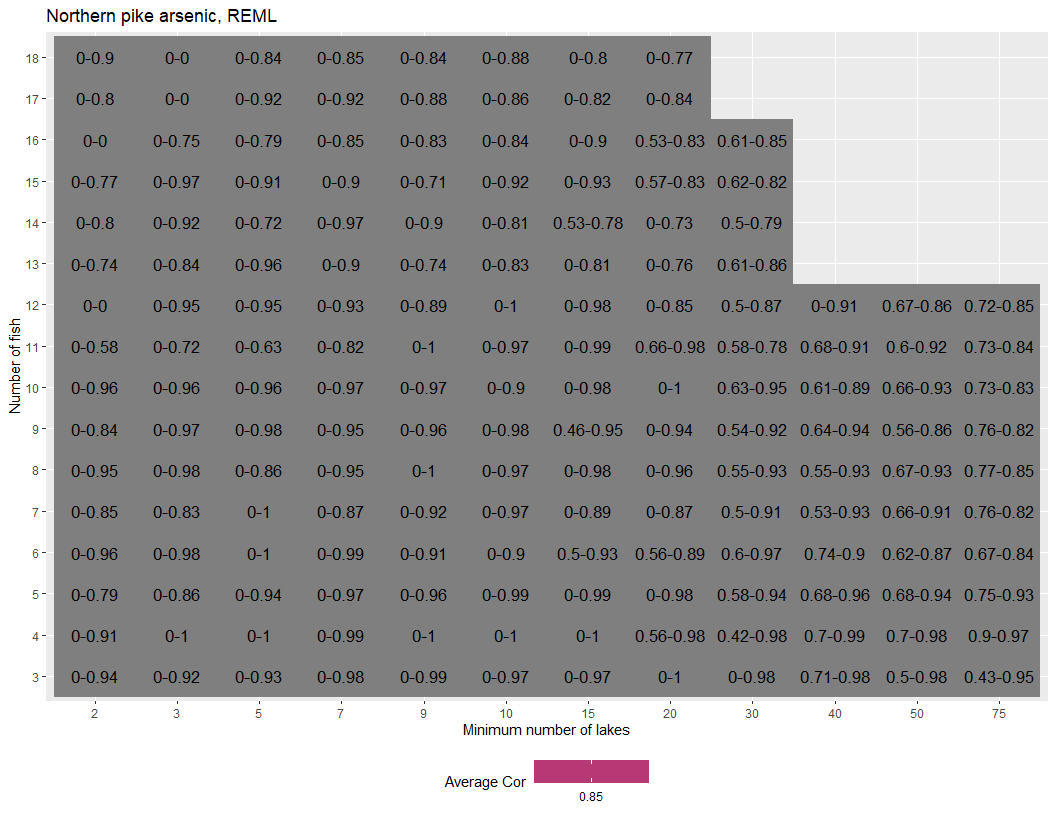


Figure 9: Range of Pearson correlation scores for simulated samplings of northern pike arsenic REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

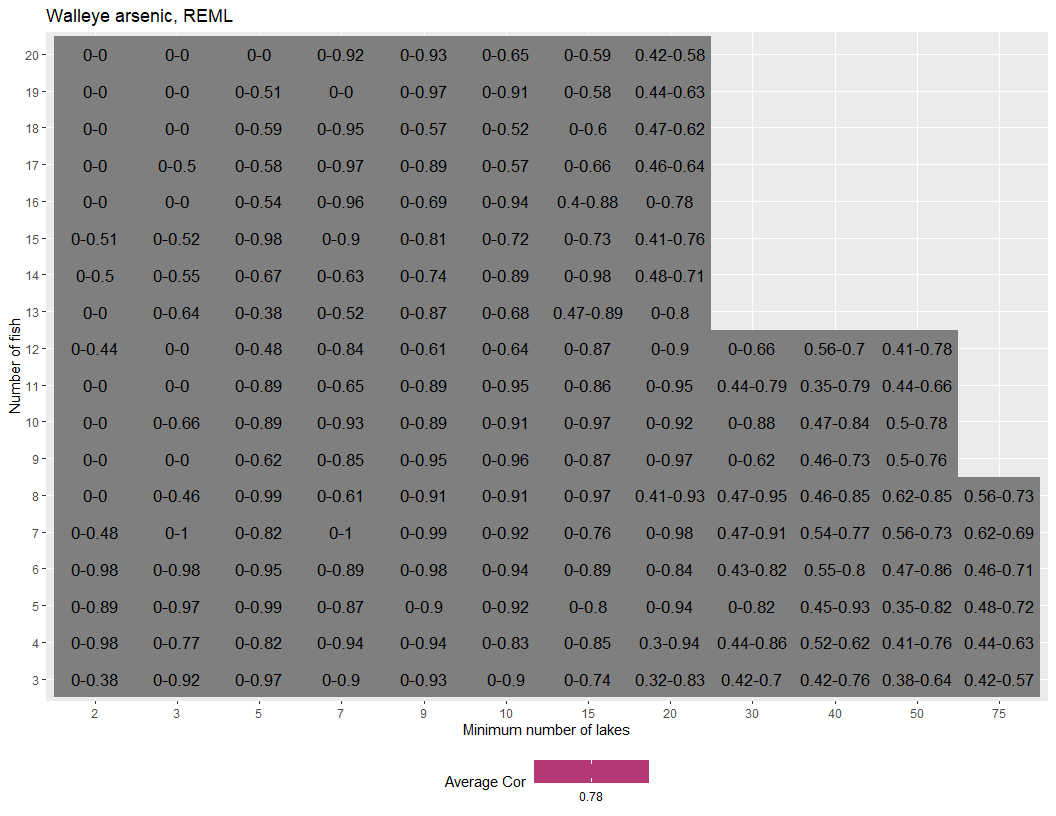


Figure 10: Range of Pearson correlation scores for simulated samplings of walleye arsenic REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where perfromance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

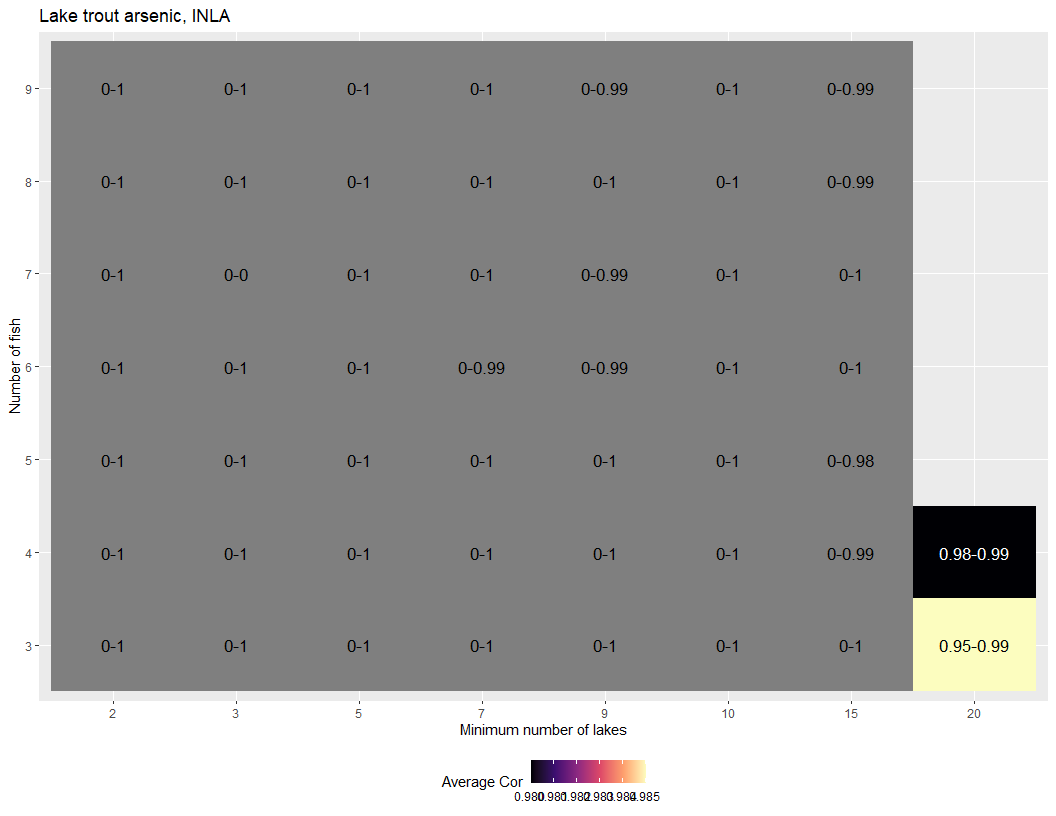


Figure 11: Range of Pearson correlation scores for simulated samplings of lake trout arsenic INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

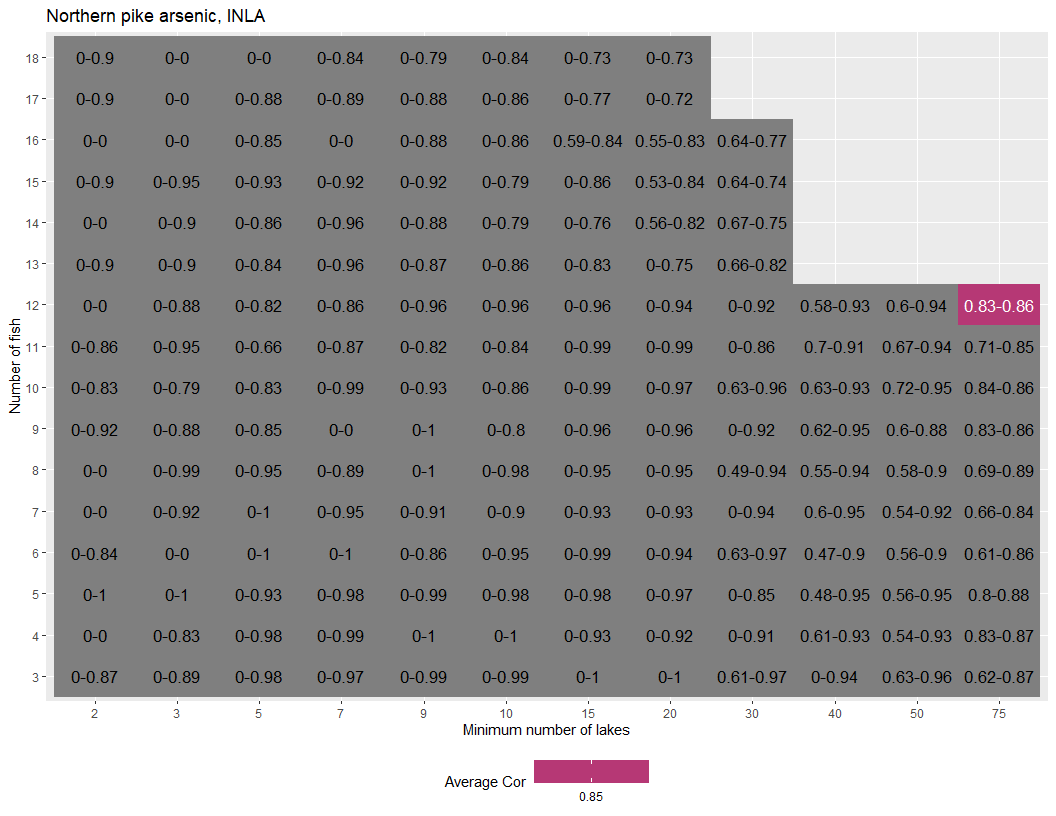


Figure 12: Range of Pearson correlation scores for simulated samplings of northern pike arsenic INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

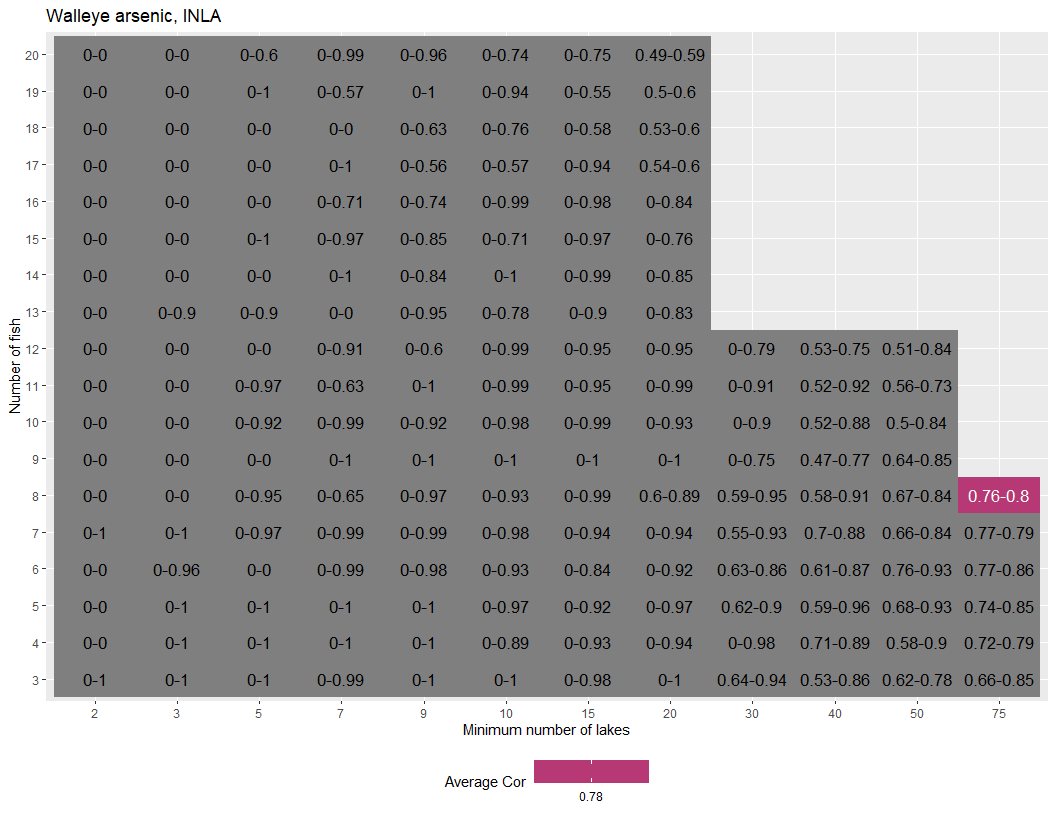


Figure 13: Range of Pearson correlation scores for simulated samplings of walleye arsenic INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where perfromance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

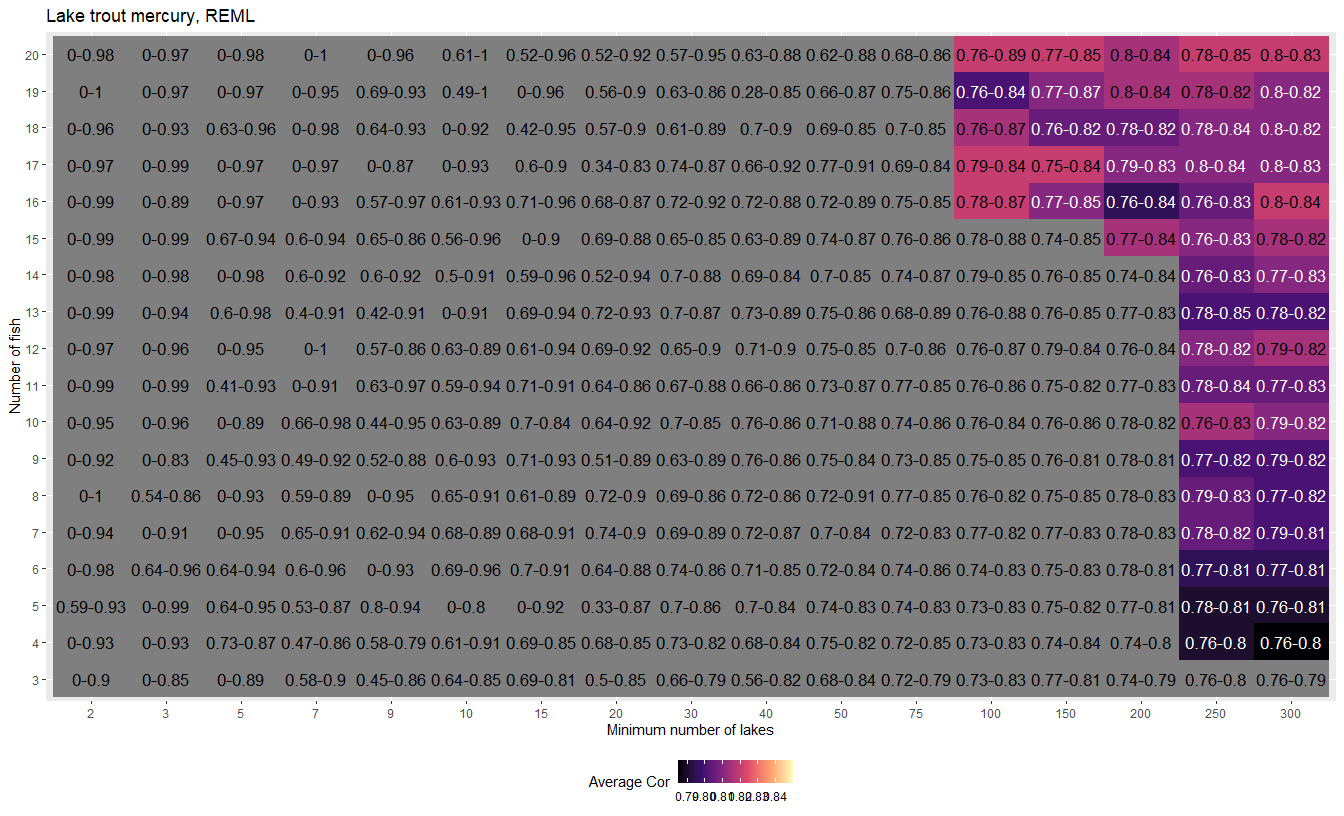


Figure 14: Range of Pearson correlation scores for simulated samplings of lake trout mercury REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

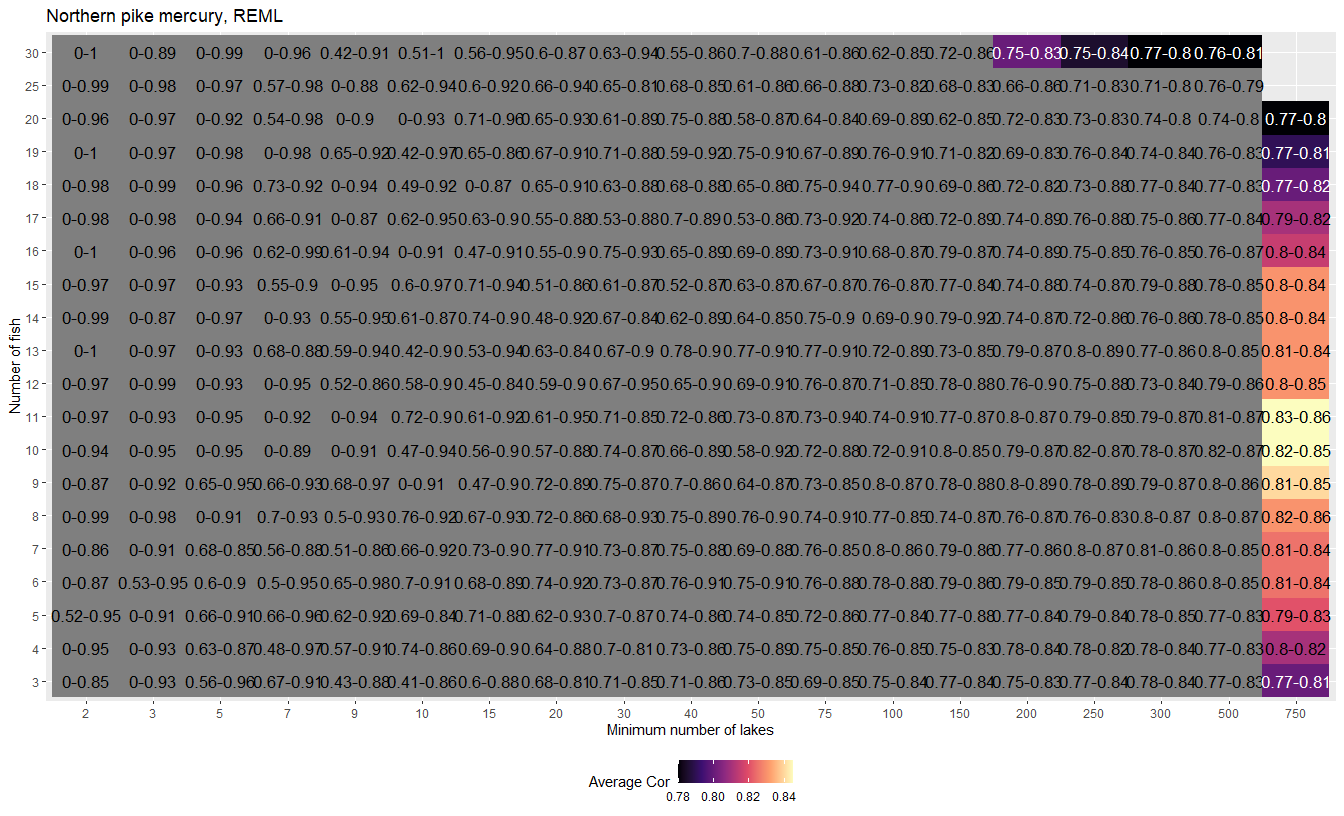


Figure 15: Range of Pearson correlation scores for simulated samplings of northern pike mercury REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

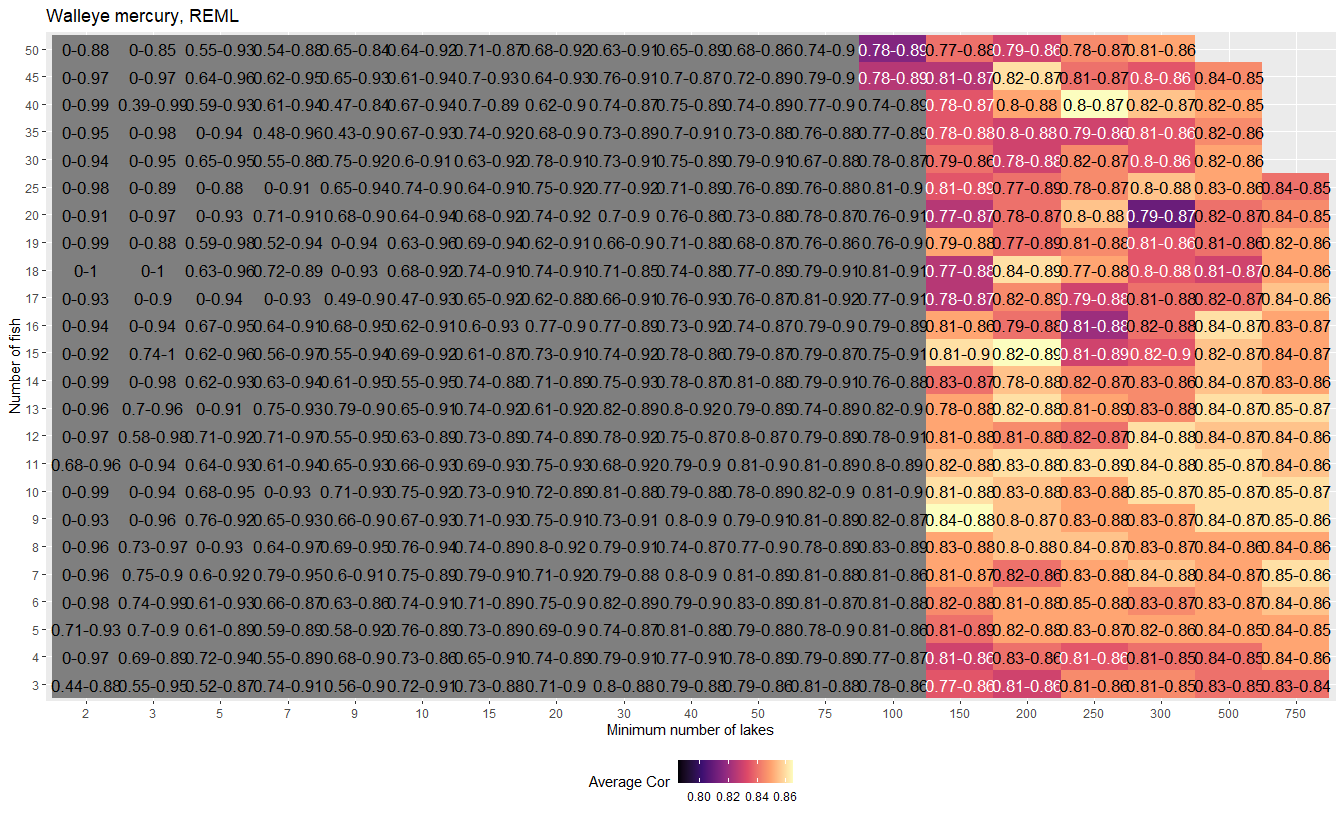


Figure 16: Range of Pearson correlation scores for simulated samplings of walleye mercury REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where perfromance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

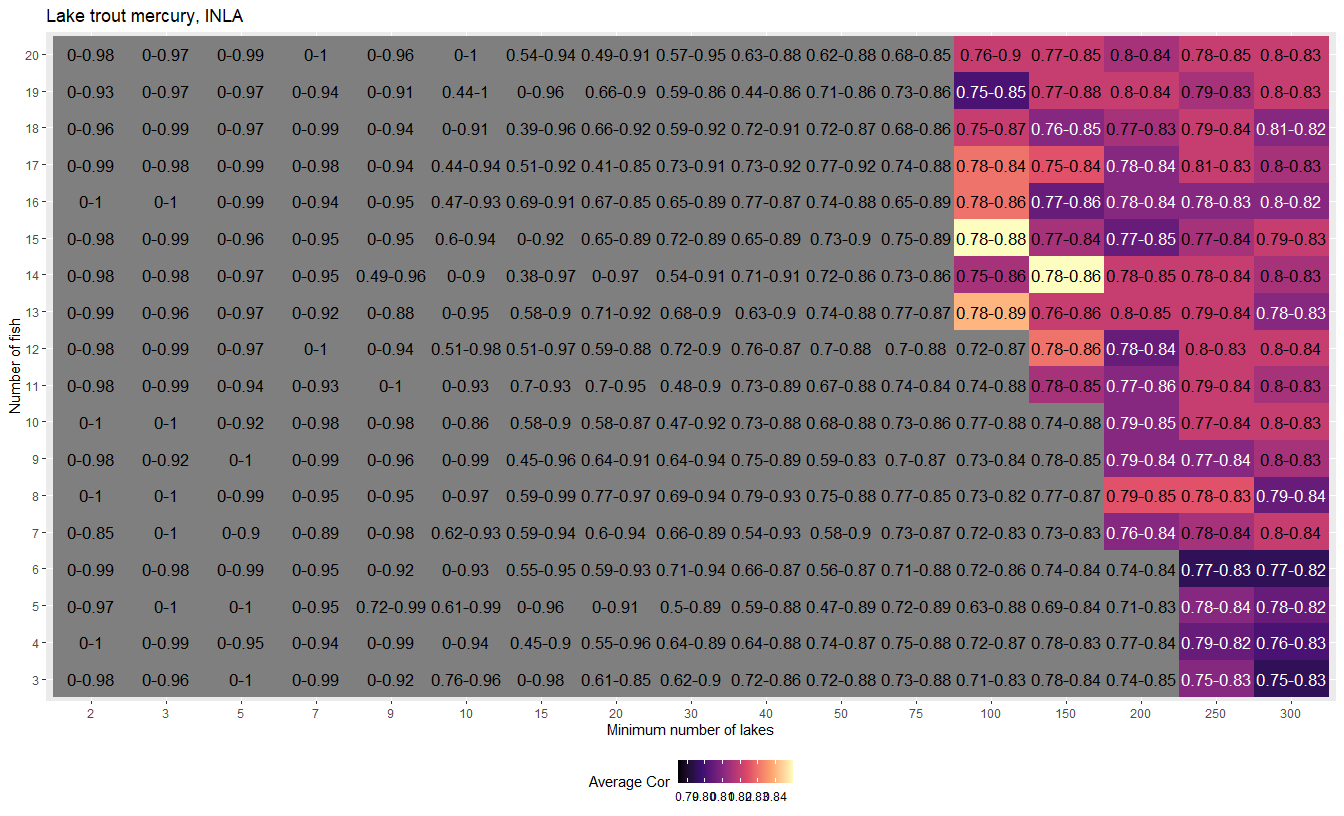


Figure 17: Range of Pearson correlation scores for simulated samplings of lake trout mercury INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

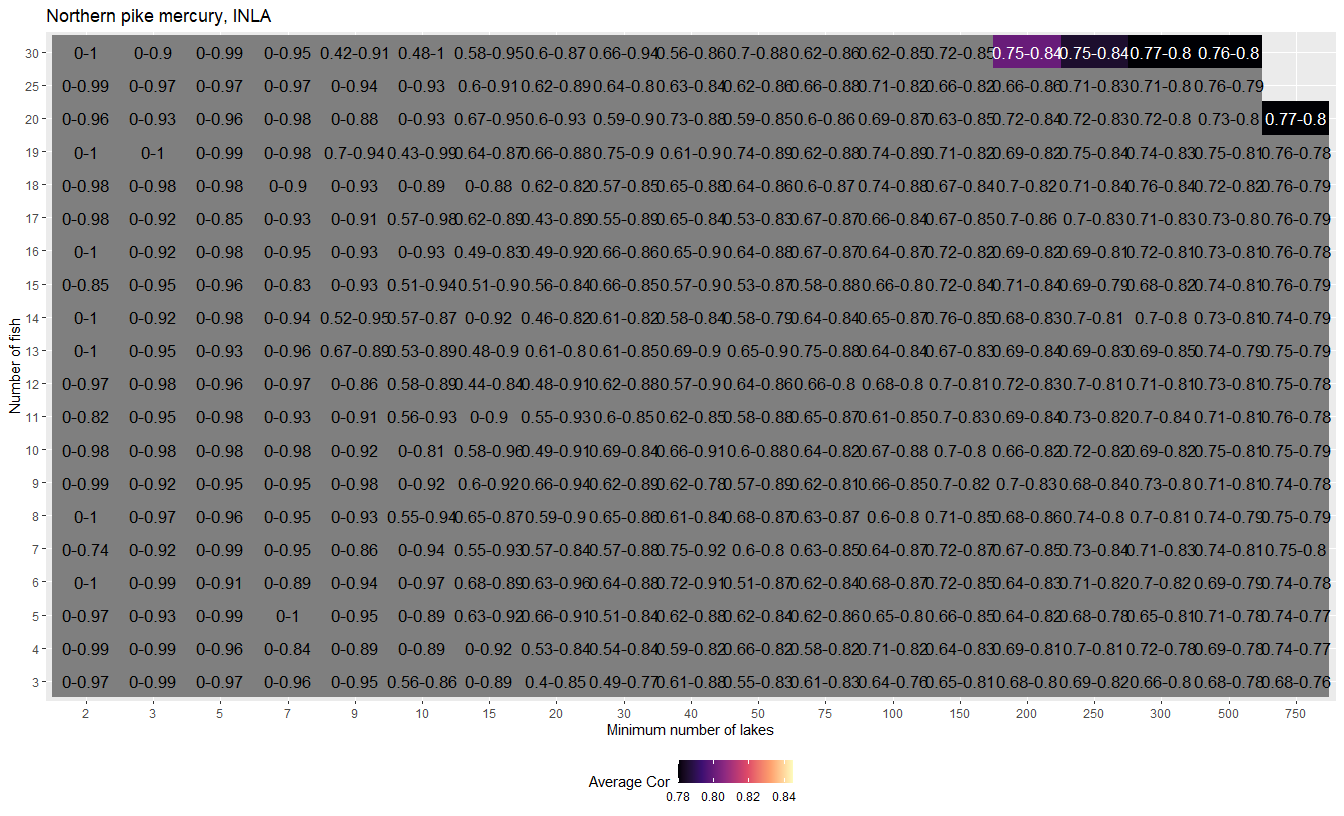


Figure 18: Range of Pearson correlation scores for simulated samplings of northern pike mercury INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

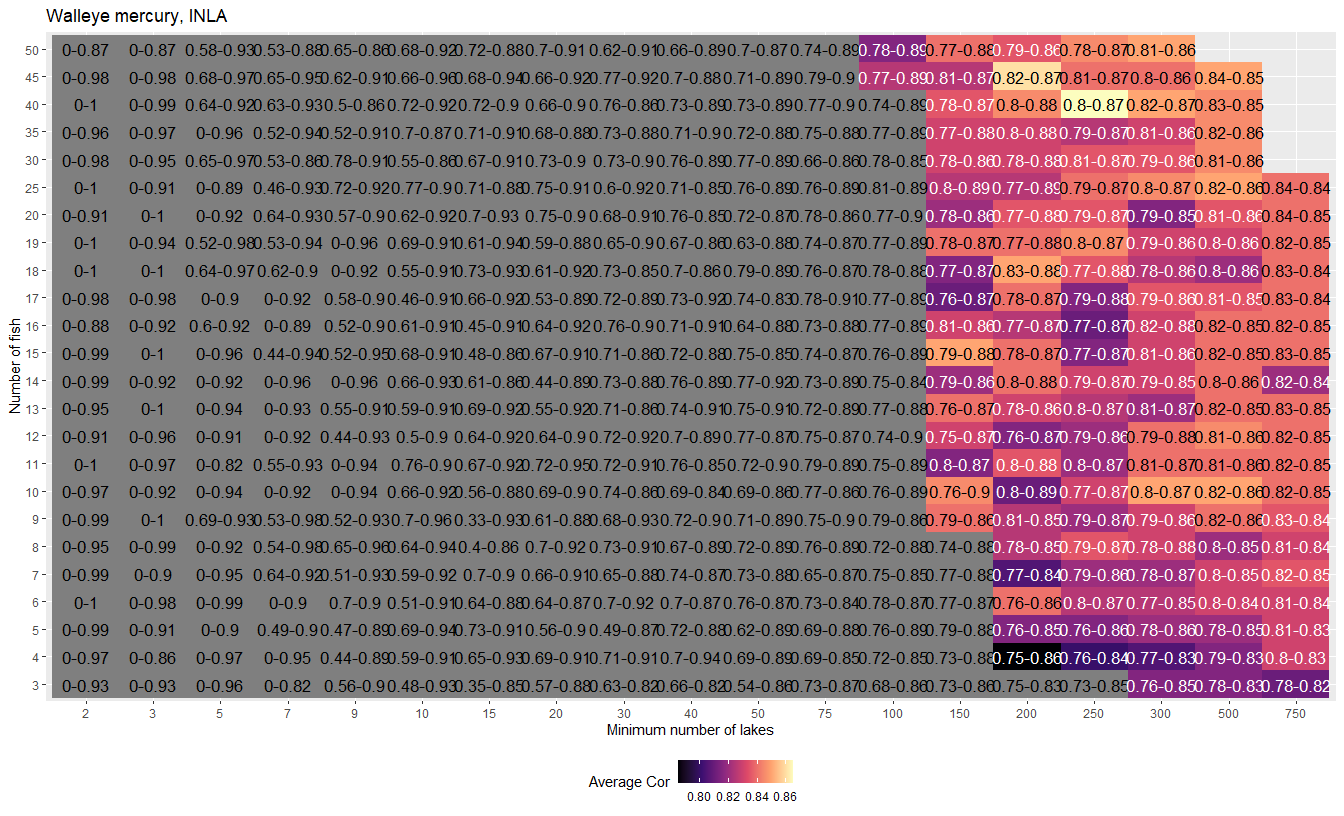


Figure 19: Range of Pearson correlation scores for simulated samplings of walleye mercury INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where perfromance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

# 

# References

(1) Ministry of Environment Conservation and Parks. Guide to Eating Ontario Fish: Advisory Database, 2023.

(2) Johnston, T. A.; Lescord, G. L.; Quesnel, M.; Savage, P.-L.; Gunn, J. M.; Kidd, K. A. Age, Body Size, Growth and Dietary Habits: What Are the Key Factors Driving Individual Variability in Mercury of Lacustrine Fishes in Northern Temperate Lakes? *Environmental Research* **2022**, 113740–113740. <https://doi.org/10.1016/j.envres.2022.113740>.

(3) Lescord, G. L.; Johnston, T. A.; Johnston, T. A.; Heerschap, M. J.; Keller, W.; Southee, F. M.; F. Meg Southee; O’Connor, C. M.; Constance M. O’Connor; O’Connor, C. M.; Dyer, R. D.; Branfireun, B. A.; Gunn, J. M. Arsenic, Chromium, and Other Elements of Concern in Fish from Remote Boreal Lakes and Rivers: Drivers of Variation and Implications for Subsistence Consumption. *Environmental Pollution* **2020**, *259*, 113878. <https://doi.org/10.1016/j.envpol.2019.113878>.

(4) Kidd, K. A.; Muir, D. C. G.; Evans, M. S.; Wang, X.; Whittle, M.; Whittle, M.; Swanson, H. K.; Johnston, T. A.; Guildford, S. J. Biomagnification of Mercury Through Lake Trout (Salvelinus Namaycush) Food Webs of Lakes with Different Physical, Chemical and Biological Characteristics. *Science of The Total Environment* **2012**, *438*, 135–143. <https://doi.org/10.1016/j.scitotenv.2012.08.057>.

(5) Ahonen, S. A.; Hayden, B.; Leppänen, J. J.; Kahilainen, K. K. Climate and Productivity Affect Total Mercury Concentration and Bioaccumulation Rate of Fish Along a Spatial Gradient of Subarctic Lakes. *Science of The Total Environment* **2018**, *637–638*, 1586–1596. <https://doi.org/10.1016/j.scitotenv.2018.04.436>.

(6) Lescord, G. L.; Johnston, T. A.; Johnston, T. A.; Branfireun, B. A.; Gunn, J. M. Mercury Bioaccumulation in Relation to Changing Physicochemical and Ecological Factors Across a Large and Undisturbed Boreal Watershed. *Canadian Journal of Fisheries and Aquatic Sciences* **2019**, *76* (12), 2165–2175. <https://doi.org/10.1139/cjfas-2018-0465>.

(7) Sumner, A. W.; Johnston, T. A.; Johnston, T. A.; Gretchen L. Lescord; Lescord, G. L.; Branfireun, B. A.; Gunn, J. M. Mercury Bioaccumulation in Lacustrine Fish Populations Along a Climatic Gradient in Northern Ontario, Canada. *Ecosystems* **2020**, *23* (6), 1206–1226. <https://doi.org/10.1007/s10021-019-00464-9>.

(8) Sonesten, L. Fish Mercury Levels in Lakes–Adjusting for Hg and Fish-Size Covariation. *Environmental Pollution* **2003**, *125* (2), 255–265. <https://doi.org/10.1016/s0269-7491(03)00051-4>.

(9) Gandhi, N.; Bhavsar, S. P.; Tang, R. W. K.; Arhonditsis, G. B. Projecting Fish Mercury Levels in the Province of Ontario, Canada and the Implications for Fish and Human Health. *Environmental Science & Technology* **2015**, *49* (24), 14494–14502. <https://doi.org/10.1021/acs.est.5b03943>.

(10) Weber, M. J.; Cashatt, D. Use of Occupancy Modelling to Identify Sample Sizes and Waterbodies with Fish Exceeding Mercury Consumption Advisory Thresholds. *Ecotoxicology* **2021**, *30* (2), 323–330. <https://doi.org/10.1007/s10646-020-02326-w>.

(11) Drouillard, K. G.; Gandhi, N.; Bhavsar, S. P.; Gewurtz, S. B.; Arhonditsis, G. B.; Petro, S. Is It Appropriate to Composite Fish Samples for Mercury Trend Monitoring and Consumption Advisories? *Environment International* **2016**, *88*, 80–85. <https://doi.org/10.1016/j.envint.2015.11.013>.

(12) Bolker, B. M.; Brooks, M. E.; Clark, C. J.; Geange, S. W.; Poulsen, J. R.; Stevens, M. H. H.; White, J.-S. S. Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution. *Trends in Ecology & Evolution* **2009**, *24* (3), 127–135. <https://doi.org/10.1016/j.tree.2008.10.008>.

(13) Zuur, A. F.; Ieno, E. N.; Walker, N. J.; Saveliev, A. A.; Smith, G. M. Dealing with Heterogeneity. In *Mixed effects models and extensions in ecology with R*; Springer New York: New York, NY, 2009; pp 71–100. <https://doi.org/10.1007/978-0-387-87458-6_4>.

(14) Fong, Y.; Rue, H.; Wakefield, J. Bayesian Inference for Generalized Linear Mixed Models. *Biostatistics* **2010**, *11* (3), 397–412. <https://doi.org/10.1093/biostatistics/kxp053>.

(15) Gabry, J.; Goodrich, B. [*Rstanarm: Bayesian Applied Regression Modeling via Stan*](https://mc-stan.org/rstanarm/); 2024.

(16) Rue, H.; Riebler, A. I.; Sørbye, S. H.; Illian, J. B.; Simpson, D. P.; Lindgren, F. K. [Bayesian Computing with INLA: A Review](http://arxiv.org/abs/1604.00860). *Annual Reviews of Statistics and Its Applications* **2017**, *4* (March), 395–421.

(17) Gómez-Rubio, V. *Bayesian Inference with INLA*; Chapman and Hall/CRC: Boca Raton, Florida, 2020.

(18) Government of Ontario. Fish Contaminants Monitoring and Surveillance Data, 2023.

(19) Kluke, C.; Lescord, G.; Johnston, T. A.; Kielstra, B.; Lock, A.; Bhavsar, S. P.; Gunn, J. Spatial Patterns and Environmental Factors Related to Arsenic Bioaccumulation in Boreal Freshwater Fish. *Canadian Journal of Fisheries and Aquatic Sciences* **2022**. <https://doi.org/10.1139/cjfas-2022-0106>.

(20) Ministry of Environment Conservation and Parks. Guide to Eating Ontario Fish. *ontario.ca*, 2023.

(21) Eagles-Smith, C. A.; Ackerman, J. T.; Willacker, J. J.; Tate, M. T.; Lutz, M. A.; Fleck, J. A.; Stewart, A. R.; Wiener, J. G.; James G. Wiener; Evers, D. C.; Lepak, J. M.; Davis, J. A.; Pritz, C. F. Spatial and Temporal Patterns of Mercury Concentrations in Freshwater Fish Across the Western United States and Canada. *Science of The Total Environment* **2016**, *568*, 1171–1184. <https://doi.org/10.1016/j.scitotenv.2016.03.229>.

(22) Eckley, C.; Eagles-Smith, C.; Luxton, T.; Hoffman, J.; Janssen, S. E. Using Mercury Stable Isotope Fractionation to Identify the Contribution of Historical Mercury Mining Sources Present in Downstream Water, Sediment and Fish. *Frontiers in Environmental Chemistry* **2023**. <https://doi.org/10.3389/fenvc.2023.1096199>.

(23) Negrazis, L.; Kidd, K. A.; Erdozain, M.; Emilson, E. J. S.; Mitchell, C. P. J.; Gray, M. A. Effects of Forest Management on Mercury Bioaccumulation and Biomagnification Along the River Continuum. *Environmental Pollution* **2022**, 119810–119810. <https://doi.org/10.1016/j.envpol.2022.119810>.

(24) Lepage, A. T.; Lescord, G. L.; Lock, A.; Johnston, T. A.; Gandhi, J.; Gunn, J. M. Biodilution of Organic Species of Arsenic in Freshwater Food Webs. *Environmental Toxicology and Chemistry* **2024**, etc.5817. <https://doi.org/10.1002/etc.5817>.

(25) Bates, D.; Maechler, M.; Bolker, B.; Walker, S. *Lme4: Linear Mixed-Effects Models Using Eigen and S4*; 2020.

(26) Rue, H.; Lindgren, F.; Teixeira Krainski, E. *INLA: Full Bayesian Analysis of Latent Gaussian Models Using Integrated Nested Laplace Approximations*; 2024.

(27) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023.

(28) Wickham, H. *Tidyverse: Easily Install and Load the Tidyverse*; 2023.

(29) Wickham, H.; Chang, W.; Henry, L.; Pedersen, T. L.; Takahashi, K.; Wilke, C.; Woo, K.; Yutani, H.; Dunnington, D. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*; 2023.

(30) Kassambara, A. *Ggpubr: Ggplot2 Based Publication Ready Plots*; 2023.

(31) Somers, K. M.; Jackson, D. A. Adjusting Mercury Concentration for Fish-Size Covariation: A Multivariate Alternative to Bivariate Regression. *Canadian Journal of Fisheries and Aquatic Sciences* **1993**, *50* (11), 2388–2396. <https://doi.org/10.1139/f93-263>.

(32) Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with Discussion). *Journal of the Royal Statistical Society B* **2009**, *71*, 319–392.

(33) Martins, T. G.; Simpson, D.; Lindgren, F.; Rue, H. Bayesian Computing with INLA: New Features. *Computational Statistics and Data Analysis* **2013**, *67*, 68–83.

(34) Visha, A.; Gandhi, N.; Bhavsar, S. P.; Arhonditsis, G. B. A Bayesian Assessment of the Mercury and PCB Temporal Trends in Lake Trout (Salvelinus Namaycush) and Walleye (Sander Vitreus) from Lake Ontario, Ontario, Canada. *Ecotoxicology and Environmental Safety* **2015**, *117*, 174–186. <https://doi.org/10.1016/j.ecoenv.2015.03.022>.

(35) Visha, A.; Gandhi, N.; Bhavsar, S. P.; Arhonditsis, G. B. Assessing Mercury Contamination Patterns of Fish Communities in the Laurentian Great Lakes: A Bayesian Perspective. *Environmental Pollution (Barking, Essex: 1987)* **2018**, *243* (Pt A), 777–789. <https://doi.org/10.1016/j.envpol.2018.07.070>.