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, Great Lakes Forestry Centre, Natural Resources Canada, [bkielstra@ecometrix.ca](mailto:bkielstra@ecometrix.ca)

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

Thomas A. Johnston, Vale Living with Lakes Centre, Ontario Ministry of Natural Resources, [tom.johnston@ontario.ca](mailto:tom.johnston@ontario.ca)

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

Robert Mackereth, Centre for Northern Forest Ecosystem Research, Ontario Ministry of Natural Resources and Forestry, [rob.mackereth@ontario.ca](mailto:rob.mackereth@ontario.ca)

Stephanie Melles, Department of Chemistry and Biology, Urban Water Research Centre, 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)

2024-09-24

1 Abstract

Concentrations of bioaccumulative contaminants in fish increase with their size and age, thus, research and monitoring of these contaminants in fish across space and time can be confounded by size covariation. To account for this, size-standardization of contaminant concentrations within fish samples is a common practice. Standardized concentrations are often estimated using within-sample regression models, also known as power series regression (referred to here as sampling event regressions, or SERs). This approach requires higher sample sizes than mixed effect models (MEMs), which suited for this application, but are not as commonly used. Herein we compare SERs to three MEM 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 MEM approaches generated size-standardized concentrations for small populations (e.g., <5 fish) and/or populations that lacked the range of sizes required for SER estimates, with comparable residual and RMSE to 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 across a range of scenarios with sample-size limitations. Additionally, we provided example code for prediction using the R-INLA package to enable 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 which are assimilated faster than they can be excreted (e.g., mercury, organochlorides). Thus, 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. Standardization is also used to control for body size (or age) covariation when comparing contaminant concentrations among study sites (e.g., investigations into drivers of contaminant concentrations due to land use, forest change, and natural variation2–5) or among different time periods (e.g., temporal trend analysis).

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 (often times, these relationships are power series regressions), and this equation is used to estimate the contaminant concentration at a chosen representative size (e.g., 500 g or 1000 g) for each fish population1,4,6–8. This approach (referred to here as Sampling Event Regressions, or SERs) 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, with fewer sample size limitations, are mixed effects modelling approaches. These approaches offer the advantage of ‘borrowing’ information from other sampling events across an entire dataset for to estimate overall “average” relationships while accounting for deviation in absolute values (i.e., random intercepts) or relationships (i.e., random slopes) among explanatory variables (e.g., waterbodies, times) to generate sample-level predictions and confidence intervals of contaminant concentrations for a given fish size12,13. By borrowing strength from other observations this approach can potentially increase accuracy and allow for estimates even when sample sizes are low, or when there is an inadequate range of body sizes for the SER approach.

A likely barrier to wide use of these mixed model methods, however, is the plethora of possible approaches, all of which have varied complexity, accuracy, and computational requirements. The most accessible implementation is the restricted maximum likelihood approach (REML), but 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 that facilitate the inclusion of confidence intervals around estimates are available15,16. This includes 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: MCMC uses randomized resampling of both samples and model parameters16, whearas INLA using the (faster) Laplace approximations for probability distributions for each model parameter15,17.

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) approach. 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, and that mixed models would still allow for 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 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 containing total Hg and As concentrations ([Hg] and [As] respectively) in fish muscle and, fish attribute data (lengths, weights, species) from the Ontario Ministry of Environment, Conservation, and Parks (MECP)18 and from data collected in Northern Ontario for Arsenic studies3. Although individual fish sampling protocols varied, muscle tissue sent to MECP was 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. Arsenic data collected as part of the Lescord et al3 As study were generated at the ISO 17025 accredited Biotron trace-metal laboratory at the University of Western Ontario using EPA method 3052 and 200.8 using microwave digestion and ICP-MS.

We included Hg because it is a highly bioaccumulative contaminant of global concern that is routinely monitored in fish and used to inform consumption guidelines19. Size adjustments of [Hg] are also common using regression (i.e., SERs) to account for [Hg] variation across environmental gradients in research3,7,20–22. We include As as a representative contaminant that has also been size-adjusted in research3,23, but is less bioaccumulative in fish, showing mixed or weaker relationships with metrics of body size3,23,24. We further limited data to only inland lakes (i.e., excluding the Laurentian 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, and they are commonly consumed by anglers, so accurate estimation of contaminant concentrations is important for assessing human health risk. Our final dataset included [Hg] in 37923 fish and [As] in 1001 fish. The datasets had a normal distribution of values, with fish between 400 and 1100 g wet weight being well represented in the data (Figure S4). Generally, most fish in the dataset had contaminant concentrations < 1.3 ug/gramwet weight [Hg] and < 0.4 ug/gwet weight [As] (Figure S4).

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 500 g and 1000 g wet weight, 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 S2).

4.2 Standardization Approaches

4.2.1 Sampling event regressions (SER)

For each sample of fish representing a unique contaminant-species-waterbody-year combination (i.e., a sampling event), we developed log-contaminant concentration (in ug/gwet weight) by log-wet weight (in g) regression models. SERs were limited to cases where there were a minimum of 5 sampled individuals, per Kluke et al, 202323, and also where the range of fish sizes allowed interpolative prediction (i.e.,training sets needed to include values outside of the 400-500 g or 900-1100 g test set ranges). Our data filtering may have been more lenient but also less subjective when compared to methods used in research, where researchers may curate data and remove outliers from each lake regression on a case-by case basis. 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-wet 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, υ0j is the random intercept for waterbody *j*, υ1j is the random slope for waterbody *j*, υ0k 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* v1.1.34 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)16 and the priors were the coefficients of an equivalent generalized linear mixed effects 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. For REML models the best linear unbiased prediction, whichcan be interpreted as an approximation of the mean and, for the bootstrapped REML, INLA and MCMC predictions we used the the median derived from predicted test value distributions, back transformed to concentrations using the *exp* function in R, as a comparison to measured values, and also to calculate root mean squared error (i.e., 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 individually from the training data of 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 these results to provide context about the bioaccumulation of As and how that may have effected model performance. We evaluated if the median concentration was an accurate predictor of test values to demonstrate if there was no Arsenic bioaccumulation. We did not include this comparison for [Hg], as the bioaccumulation of Hg in these species is well documented.

4.4 Comparing model fit and predictive accuracy

To assess the predictive accuracy of each approach, we compared the back-transformed predicted vs observed concentrations for individual fish from the test data set using the linear relationship of the values with the *lm* function in R and Pearson correlation with the *cor.test* function in R. We also assessed the fit of each approach by calculating the RMSE of the training data set for each sampling event (i.e., distinct year-waterbody combination). Slopes and model fit based on test data reflect the generalized (i.e., out-of-sample) predictive capability of a model, while correlation values reflect the accuracy of the model in representing the trends present in the training data.

We ran an additional analysis to assess how model performance changes with different fish and lake number sampling scenarios using each year-waterbody as distinct sample events. We trained REML and INLA models on randomly sampled training sets with a range of lake numbers and fish numbers to evaluate model predictive performance. We used the less computationally intensive INLA approach to represent both Bayesian approaches, as we determined over the course of this study, INLA and MCMC were highly similar in terms of predictive capability. This was run for a range of numbers of lakes (2,3,5,7,9,10,15,20,30,40,50,75,100,150,200,250,300,500,750,1000) and a select number of fish sample sizes for each lake (3-20, 25, 30, 35, 40, 45, 50), with 10 replicate samples for each combination. We assessed the performance of these models by evaluating whether stable model performance was achieved (i.e., maintained average and minimum Pearson correlation coefficient values of 0.8 and 0.75, respectively, as number of lakes and number of fish increased).

Computational times were assessed in the context of analyses performed in R version 4.3.1 (2023-06-16 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 the approaches tested, 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 measured values (r > 0.75) and had favourable computational speeds when compared to other approaches. Setting aside inferential differences, 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.

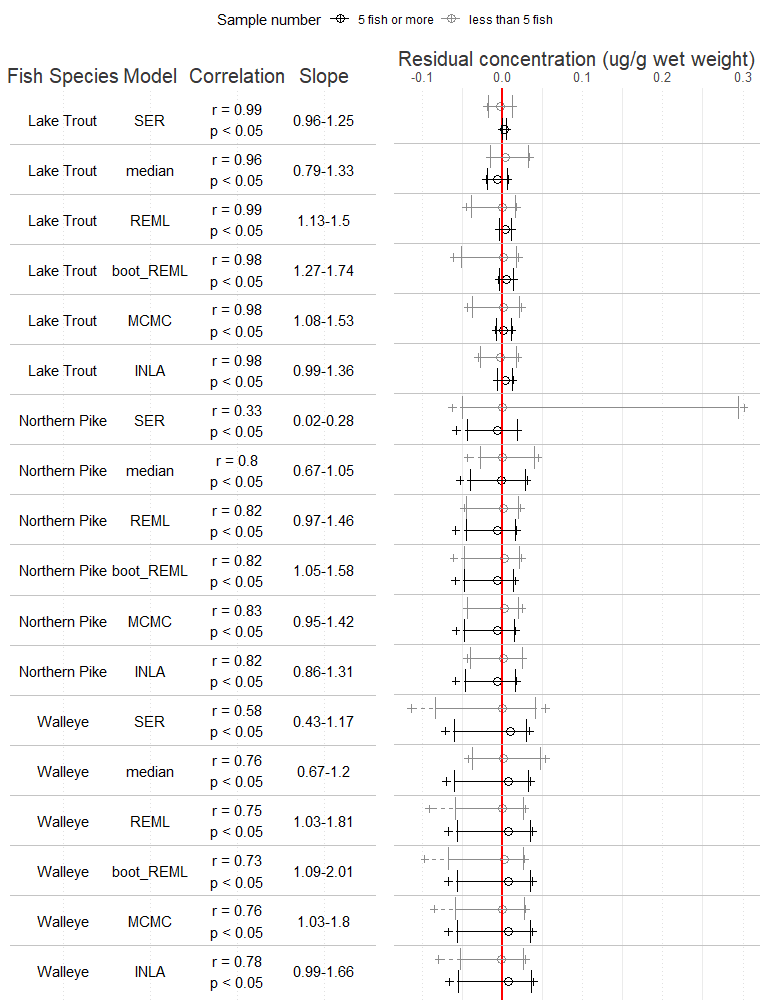


Figure 1: Comparison of predicted vs observed relationships for muscle [As] using predicted test-set values generated from six models. Predictions were made for 500g ± 100g and 1000g ± 100g fish of three species. Correlation statistics (Pearson’s r and p-value) and the 95th confidence interval of slopes from linear association of predicted and measured values are indicated for each model. For both correlations, and slopes, values closer to 1 represent stronger representations of overall trends, and actual values, respectively. Residuals are plotted 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.

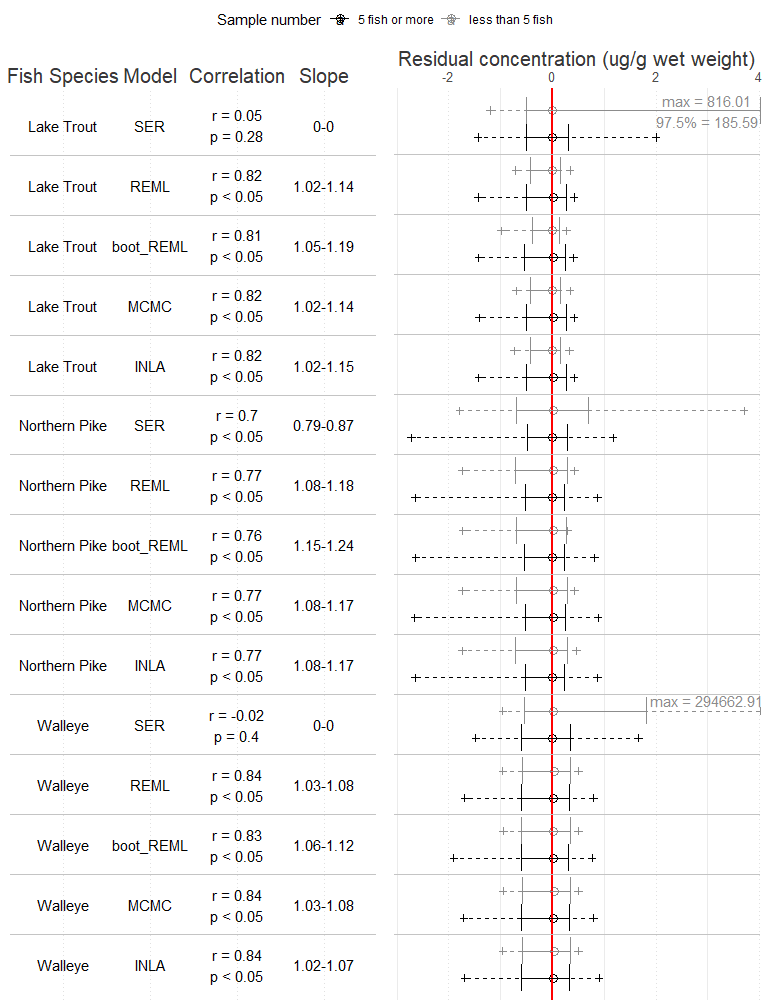


Figure 2: Comparison of predicted vs observed relationships for muscle [Hg] using predicted test-set values generated from six models. Predictions were made for 500g ± 100g and 1000g ± 100g fish of three species. Correlation statistics (Pearson’s r and p-value) and the 95th confidence interval of slopes from linear association of predicted and measured values are indicated for each model. For both correlations, and slopes, values closer to 1 represent stronger representations of overall trends, and actual values, respectively. Residuals are plotted 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 the text on the right side of the plot.

We found that mixed effects approaches performed better than SER for representing the trends in the data, as represented by improved correlations. The mixed models also enabled predictions of values for low sample sizes with only slight accuracy reductions in the As models or marked accuracy improvements in the Hg models, compared to SER ([As] Figure 1, [Hg] Figure 2, Figure S5). The REML, MCMC and INLA values were generally good at representing the trends in the samples, as they were more strongly correlated to actual concentrations (r ranging from 0.73 - 0.99 in [As] and 0.76 - 0.84 in [Hg]) than predictions from SER models (r ranging from in [As] and 0.02 - 0.7 in [Hg]) (Figure 1, 2, Figure S5). Prediction accuracy of for [Hg] with the REML, MCMC and INLA (slopes deviating 0.02 - 0.24 from 1) approaches was better than SER (slopes deviating 0.13 - 1 from 1), but [As] prediction in REML, MCMC and INLA (slopes deviating 0.01 - 1.01 from 1) was slightly poorer, compared to SER (slopes deviating 0.04 - 0.98 from 1). The reduced prediction accuracy in [As] was small, being generally with slopes within 0.05 of SER values. The reduction in [As] mixed model performance may have been due to the lower overall dataset sizes for the [As] models, and possibly, also influenced by the assumptions of relationships between weight and [As] made in the structure of the mixed models. As [As] was more accurately predicted using medians for Northern Pike and Walleye than any of the linear modelling approaches, that appears to be the case for those two species. Lake Trout might be an exception from this, as it could be prone to slightly more bioaccumulation of As due to it’s slow growth rate. The increased accuracy of [Hg] models can be attributed to the elimination of influential residuals that resulted from the magnification of error after back-transformation of the log-transformed SER predictions. Also, the residuals between the back-transformed predicted values and measured concentrations for the different mixed modeling approaches were generally similar or smaller than those of the SER approach for both [As] (Figure 1) and [Hg] (Figure 2) models. Interestingly, attempts to improve the REML approach with bootstrapping did not increase accuracy of predictions, even resulting in decreased correlation strengths and larger deviations of slopes from a one-to-one relationship in some cases (specifically for the [Hg] Northern Pike and Lake Trout models (Figure 2), and the [As] Walleye model (Figure 1) ([As]). 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 model approaches were just as effective in lakes with <5 fish as they were in lakes with larger numbers of fish. SER is limited compared to mixed model approaches as it cannot borrow information from other samples. In mixed models, an overall slope is created across all samples and individual lake and year slopes are modelled relative to the overall slope: through this, they effectively borrow information from other samples to enhance estimation. Though the mixed model approach was only required for improving prediction for the highly bioaccumulative contaminant (i.e., [Hg]), the accuracy of mixed model approaches were always comparable or better than SER or basic population medians. INLA was often the best performing mixed model approach. Mixed model approaches improved prediction for datasets with less than 6000 training fish, which would be applicable to many studies. The mixed model approaches effectively added more good or highly accurate predictions to the dataset, which may explain the increased correlation of predictions and measured values. Nevertheless, the mixed model approaches are a suitable solution for contaminant prediction in a variety of dataset sizes and contaminants of variable bioaccumulative potential.

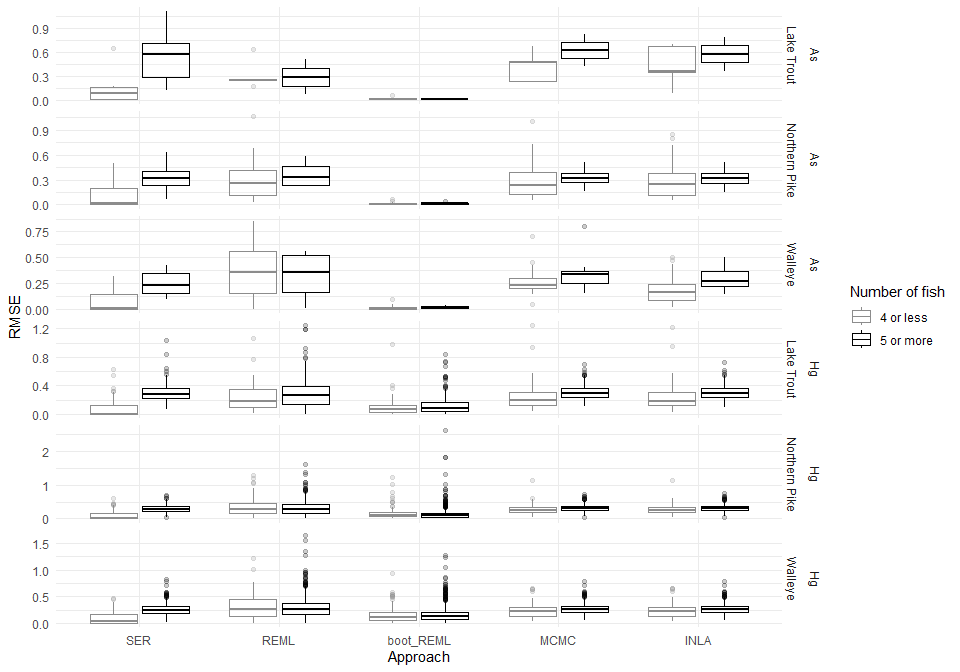


Figure 3: Root mean squared error (RMSE) from sampling events trained with different numbers of fish. RMSE of training set sampling events are compared to the number of fish used from the sampling event that were included in the training dataset. For events with less than 5 fish, and events with 5 or 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. Higher RMSE means there was more variation in the data overall, and a wider box means there was more variability in the fit of predicted training values between distinct waterbody-year sample events. 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 RMSE for training data of the different sampling events illustrated how 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 higher 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. This is likely due to bootstrapping procedures overfitting results to the training data, which reduced the extrapolative predictive capability of the models.

INLA and MCMC approaches had the best fits at low fish sample numbers, with fewer influential points than other mixed model approaches or SER. There was a trend amongst the models that low sample number events had lower RMSEs, likely because fewer values were fit and therefore overall variation was reduced. All approaches had RMSE that generally increased and stabilized at about 6 fish per sampling event but models were fairly comparable at event sizes below 6, with the exception of some events with influential points (Figure S6). These results suggest that a minimum sample size of 6 fish/sample should be collected by monitoring and research programs when size standardized contaminants are used. MCMC and INLA generally appeared to have fewer large influential RMSE values, and a more even distribution of RMSE. These approaches were not restricted to minimum sample sizes and, they had improved precision and model fit, given they could account for variation in lake-level slopes and intercepts8,31–33.

INLA and MCMC models had good model fit, without a detectable loss in accuracy, making them suitable for understanding the mechanisms of prediction. 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 provided by the INLA and MCMC approaches showed that they were better at explaining the patterns in the training data than REML, and the improved fits had little or no cost to predictive accuracy (i.e., the models are better at predicting the general population trends and individual values). Thus, if a researcher is interested in understanding the mechanisms of prediction, the Bayesian approaches would be suitable options because they provide a more thorough evaluation of the selected dataset as the generation of posterior distributions to reduces the influence of individual influential values. INLA would be recommended for smaller datasets, as there was marked improvements of the INLA approach compared to MCMC on the smaller [As] datasets.

5.1 Relative computational requirements

Table 1: Computational time (in seconds) for each of the five modeling approaches for each species and contaminant combination.

| 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, these approaches are more complex and thus, can have higher computational requirements. Computation times ranged from less than one second to 47248 seconds (13.1 hours) across modelling approaches and species. Overall, INLA models were much less computationally intensive compared to the MCMC and REML\_boot implementations (Table 1). While REML was, generally, the quickest of all modeling approaches, INLA was a close second, with computational times that were orders of magnitude lower than REML\_boot or MCMC (Table 1). The overall accuracy of the INLA approach, when compared to the REML\_boot approach, and paired with its improved speed compared to the MCMC approach suggests that INLA was the most viable approach for standardizing fish contaminant values.

In our simulated samplings, we confirmed that INLA and REML were effective at creating predictions for lakes with low sample numbers, but the approaches both required a large number of lakes. Northern Pike and Lake Trout [Hg] prediction was possible in simulations where 4 fish were used, but these predictions required more than 100 lakes (Figure S13, S15, S14, S16). For example, prediction of Lake Trout [Hg] 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 [As] datasets had very large number of lakes, which limited this line of investigation. These results show that prediction of contaminant concentrations with few fish per sample event is possible when larger, existing datasets are leveraged. We suggest that creating predictive models that leverage 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.

The simulated samplings also revealed that REML models perform better with more balanced sampling designs than INLA models. Generally, in the random samples, which had an equal number of fish selected for each lake sample event included in the sample, REML (Figure S7, S9, S11, S13, S15, S17) performed better than INLA models (Figure S8, S10, S12, S14, S16, S18). This contrasted with the results from the assessment of the test dataset, where REML and INLA models performed comparably in terms of correlation scores (Figure 1, Figure 2). As the REML models had slightly lower performance in models created from the test dataset (Figure 1, Figure 2), this suggests that the REML models are more sensitive to sampling imbalance than INLA models.

The INLA models developed from the simulated samples may have had a more consistent predictive behaviour with respect to increasing sample size (e.g., number of fish, or number of lakes) than REML models. This pattern was particularly apparent in the Walleye [Hg] dataset, where the highest average correlation scores for REML models were in lakes with 10-12 fish (Figure S17), while INLA models generally had an increase in average correlation with increases in number of fish or number of lakes (Figure S18). This characteristic of INLA models means any addition of data to the training dataset will benefit the overall INLA model, whereas the behaviour of REML models was more variable.

Each modelling approach had different benefits and drawbacks in terms of implementation. Based on the user experience during this study, the relative difficulty in application of each approach was REML<SER<MCMC<INLA. The REML approach was simple enough that a beginner R user could implement it quickly, without requiring knowledge of programming. The SER approach required more programming knowledge to implement efficiently, and a significant amount of pre-screening of data to ensure that a large enough span of fish sies were included when developing a curve. There were more documentation and reading requirements to implement the MCMC and INLA models. Additionally, INLA, with the current INLA package implementation, poses a higher risk of incorrect implementation since the model formula input deviates from the typical R notation. There is thorough documentation of the model notation in the INLA package26, 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.

Conceptually, the SER approach is the simplest to understand, given that it is essentially a group of independent linear models. However, using the SER approach to standardize the fish contaminant comparisons across lakes effectively removed some of the variation that is present in regional datasets, and thereby underestimates uncertainty. The REML, MCMC and INLA approaches account for this uncertainty in their model structure by allowing slopes and intercepts to vary by lake, and these MEM approaches quantify this variation partitioning in the error term. MEMs are more conceptually challenging in terms of model specification, however, Bayesian models have more complexity compared than the standard REML approach. The complexity of the Bayesian (MCMC), and approximated Bayesian (INLA) approaches should be considered when model tuning is required, which is often required for smaller datasets (<300 fish). Relatively few studies23,34,35 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 best practices during implementation. For researchers pursuing predictive accuracy alone, the REML approach may be the best option due to the ease of implementation. Probability distributions from Bayesian models can inform sample size adjustments: variation around intercepts can provide information on sample size and variation around slopes can indicate if the range of sites sampled are sufficient.

6 Conclusions

Mixed model approaches show promise for increasing predictive power for estimating contaminant concentrations in fish by leveraging the predictive power from 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 approach but was not able to perform as well as INLA for predicting samples with a small number of fish. REML also had poorer model fit to the training data, meaning it is less suitable for explaining the effects driving contaminant-size relationships. INLA generally was one of the less computationally intensive options, and had less RMSE in the models, making it suitable for both explanatory and predictive applications. INLA can be performed quickly, without high performance computing requirements. INLA also had comparable predictive accuracy at ~12 or fewer samples to the full dataset, which may allow for reductions in sampling effort for studies that intend to use a size-standardized contaminant concentration. Additionally, this approach allows for prediction in datasets do not conform to regression assumptions (i.e., it allows for prediction for samples that do not have homogeneity of variance, and lack balance in their sample sizes). In summary, we suggest INLA is a suitable option for size-adjustment of contaminant concentrations in fish with the potential to increase sample sizes. Use of these tools can enable greater power in investigations of important environmental drivers of contaminants in fish. A future avenue of research may be investigating the strengths of each modelling approach in different scenarios, to evaluate which models are most appropriate for prediction, and which are better at analyzing trends at various effect sizes.

7 Acknowledgements

8 Data Accessibility

The processed data, scripts used to create this manuscript, and the INLA example code 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>

9 Supplemental Tables

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

10 Supplemental Figures

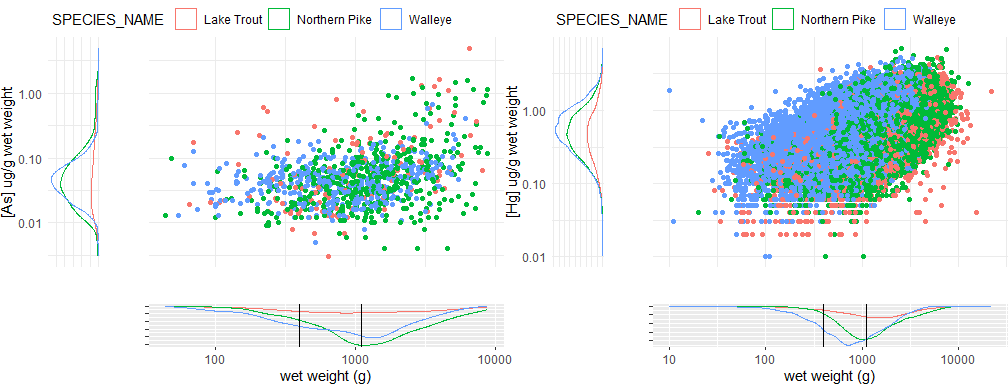


Figure 4: Number of fish at different body weights and contaminant concentrations for [As] and [Hg]. Vertical black lines show the weights that are commonly used for standardization.

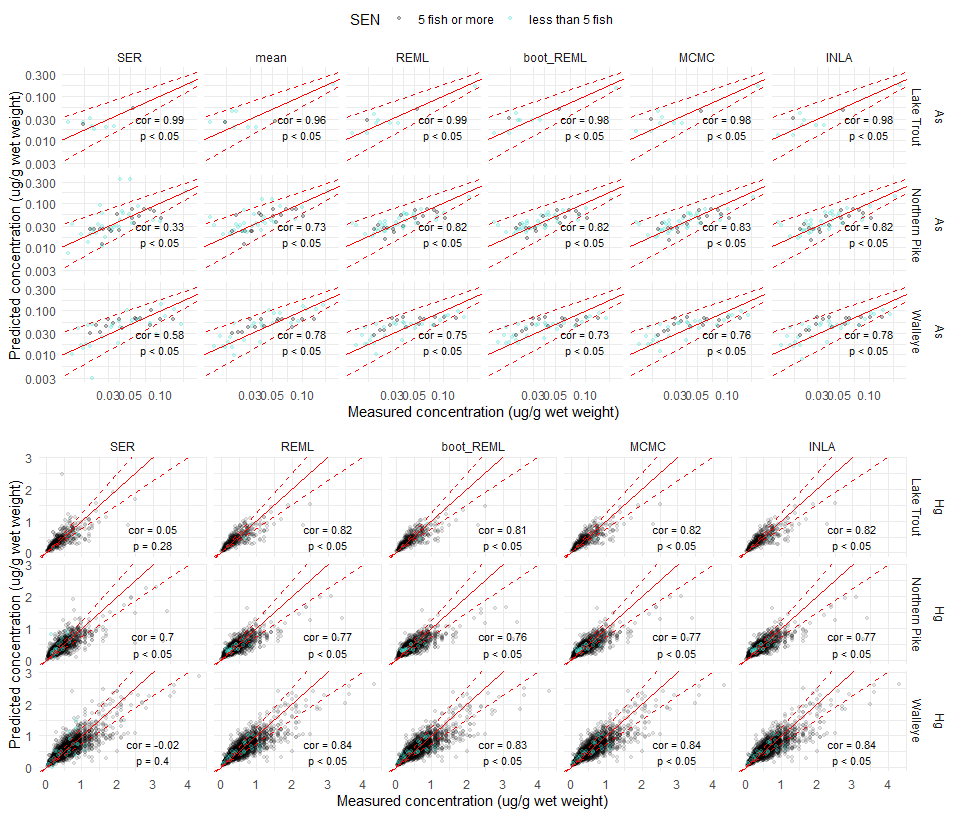


Figure 5: Predicted concentration of [As] and [Hg] in tissue of 400-600g and 900-1100g fish compared to measured values. Points display each individual fish prediction (with blue points representing predictions at sampling events with less than 5 fish, and black points representing predictions from sampling 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.

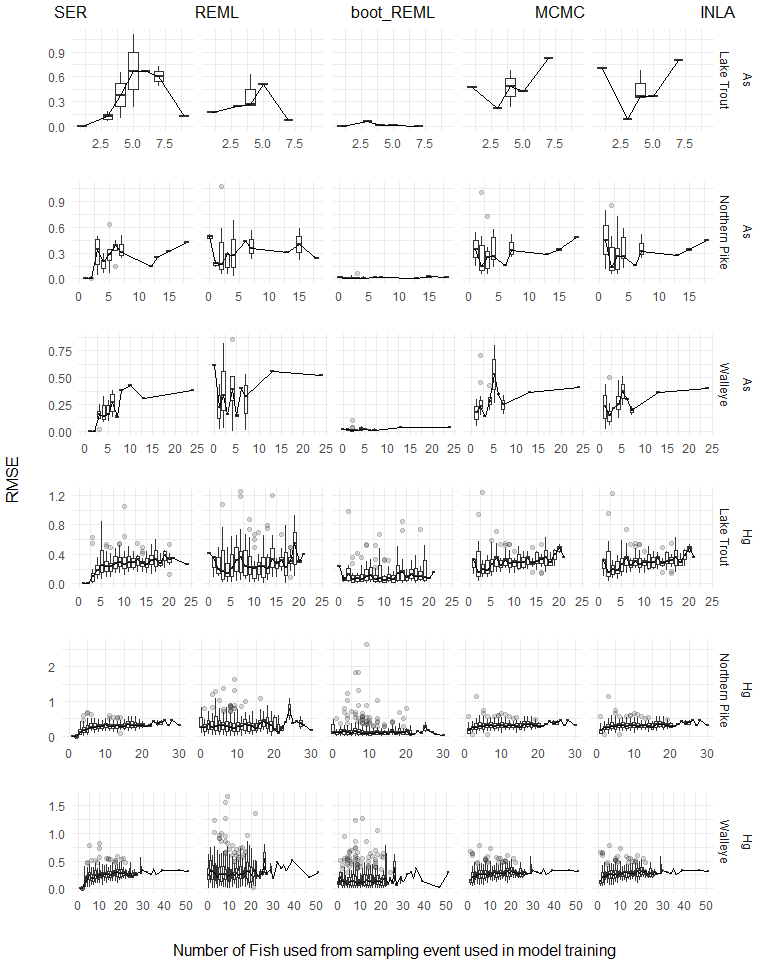


Figure 6: Root mean squared error (RMSE) of models from sampling events trained with different numbers of fish. RMSE of training set sampling events are compared to the number of fish used from the sampling event that were included in the training dataset. For each N, a boxplot displays the 25th and 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.

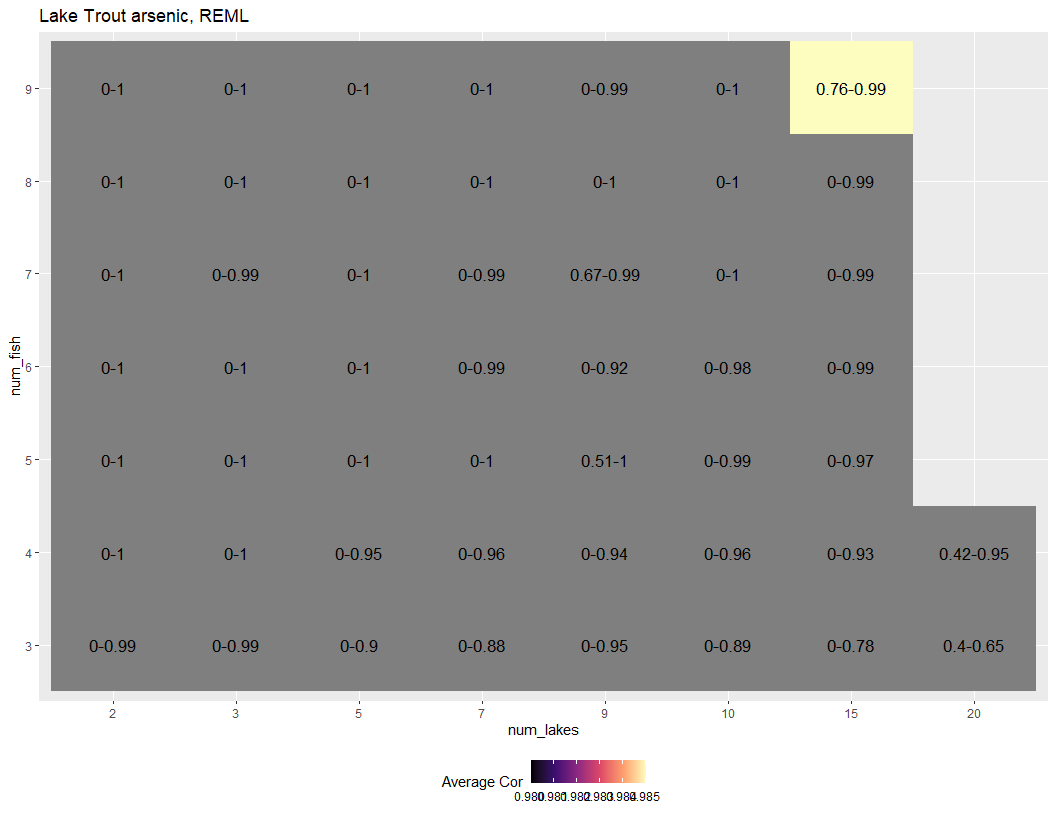


Figure 7: Range of Pearson correlation scores for simulated samplings of Lake Trout [As] REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or lake numbers are increased are indicated in grey, while remaining cells are colored with light values representing better average model performance.

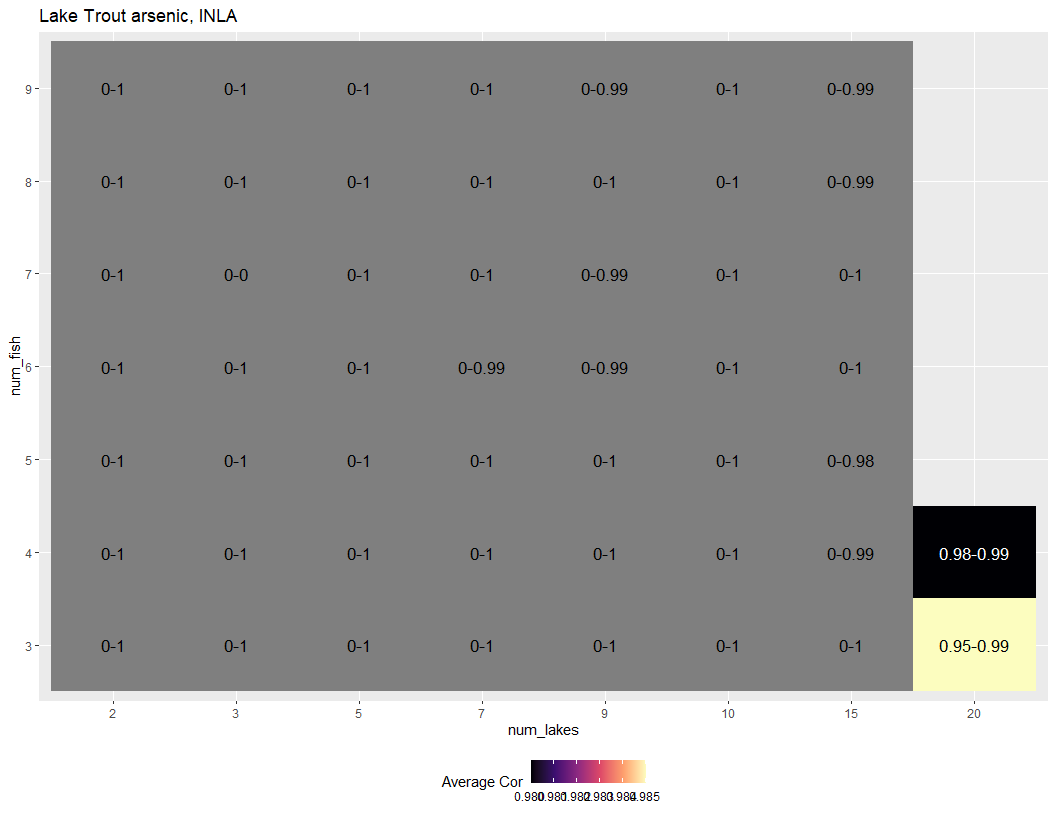


Figure 8: Range of Pearson correlation scores for simulated samplings of Lake Trout [As] INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or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

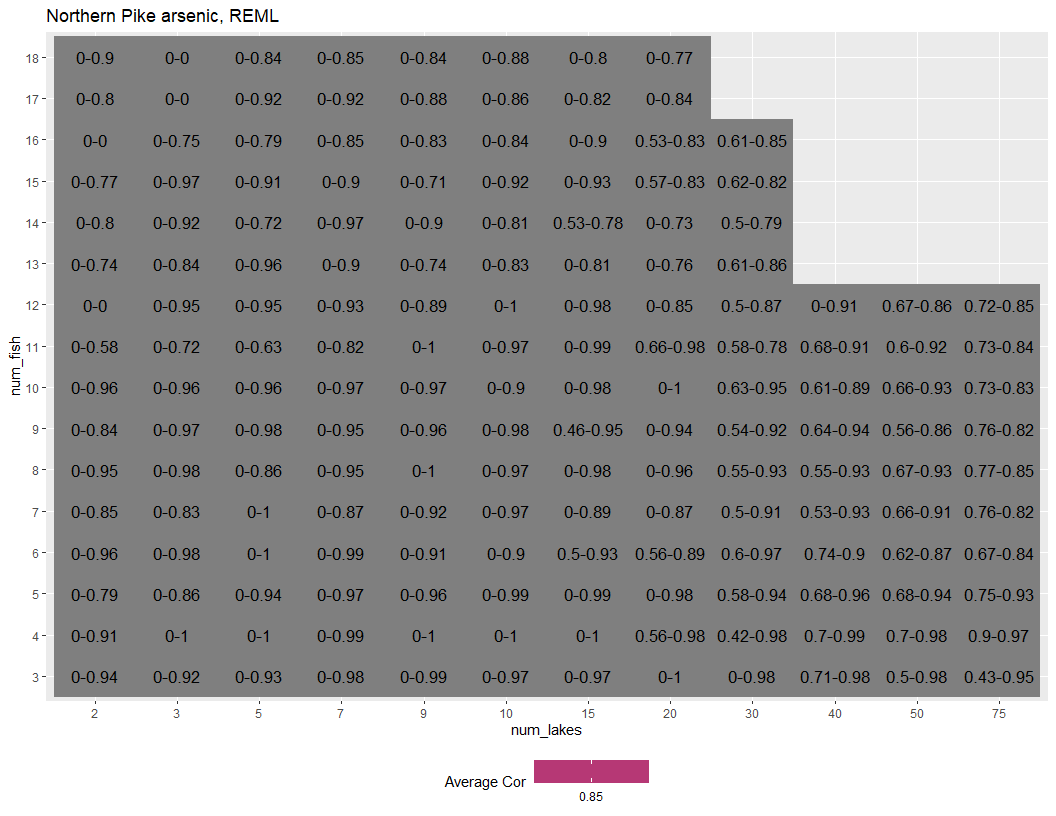


Figure 9: Range of Pearson correlation scores for simulated samplings of Northern Pike [As] REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish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

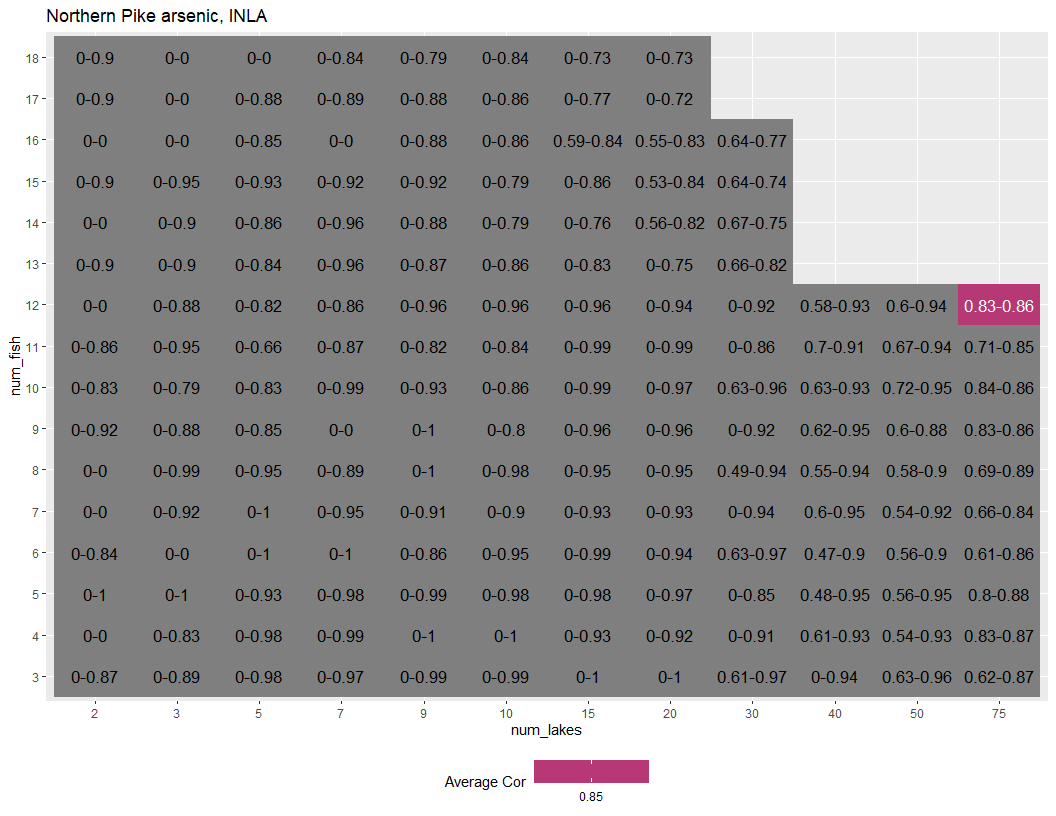


Figure 10: Range of Pearson correlation scores for simulated samplings of Northern Pike [As] INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish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

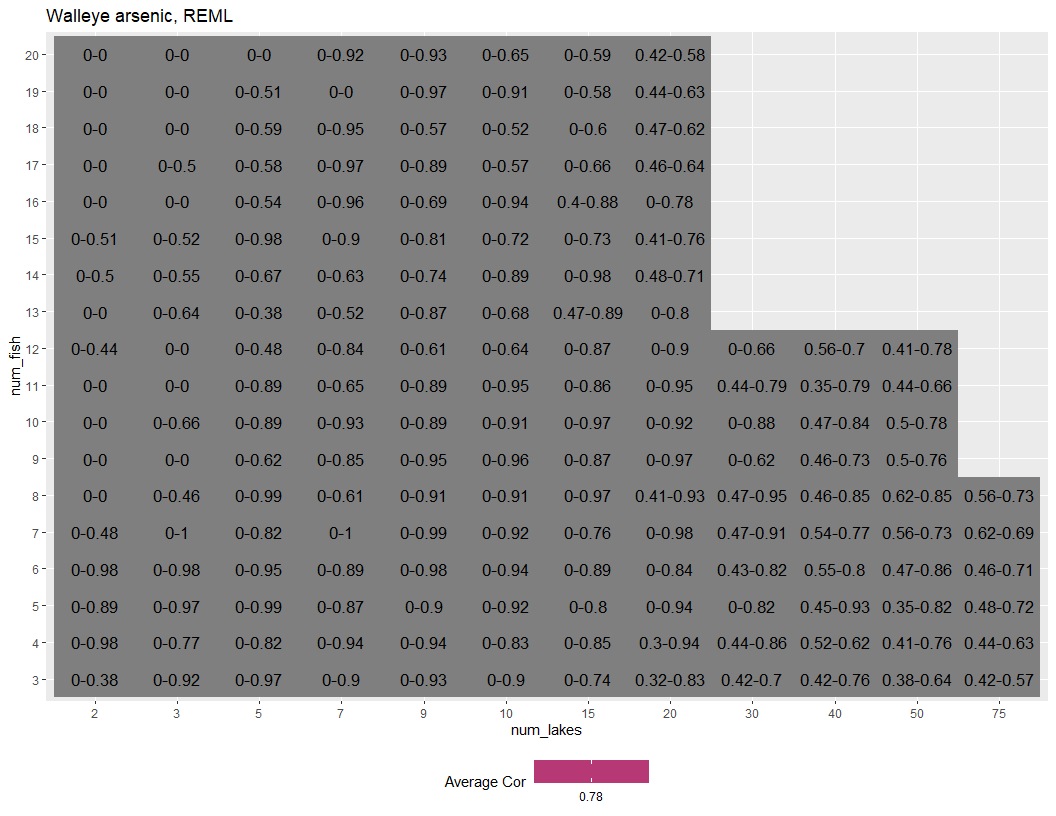


Figure 11: Range of Pearson correlation scores for simulated samplings of Walleye [As] REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or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

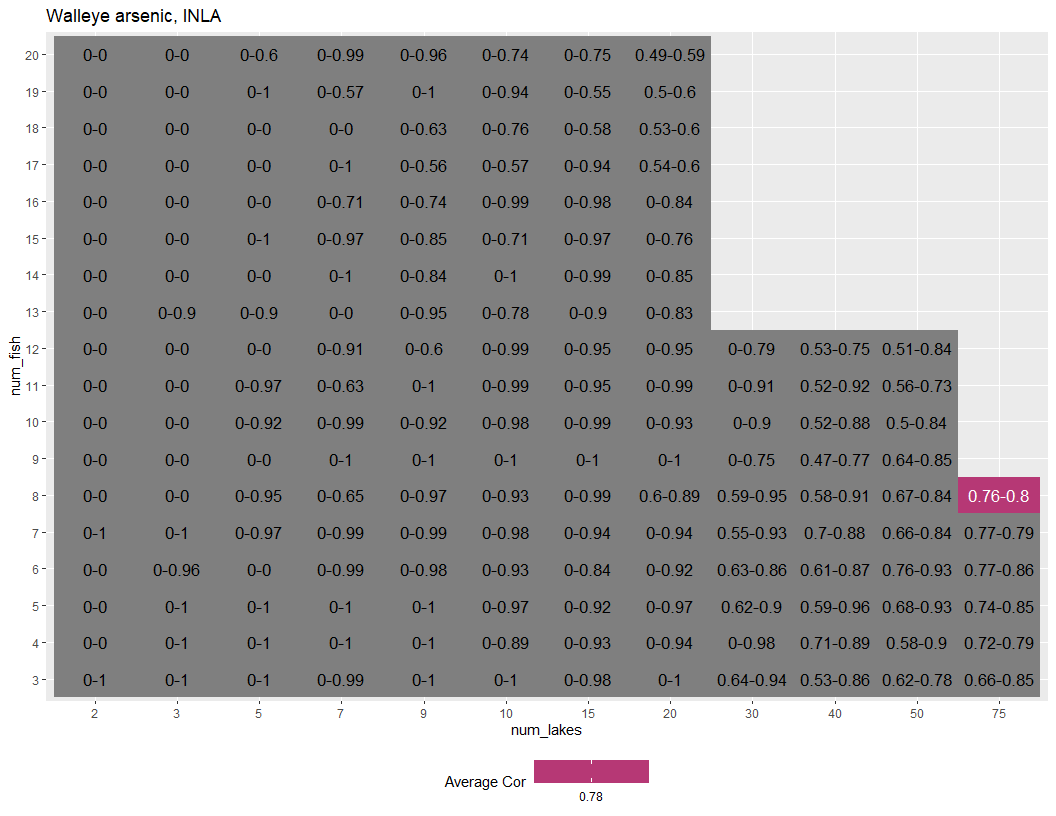


Figure 12: Range of Pearson correlation scores for simulated samplings of Walleye [As] INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or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

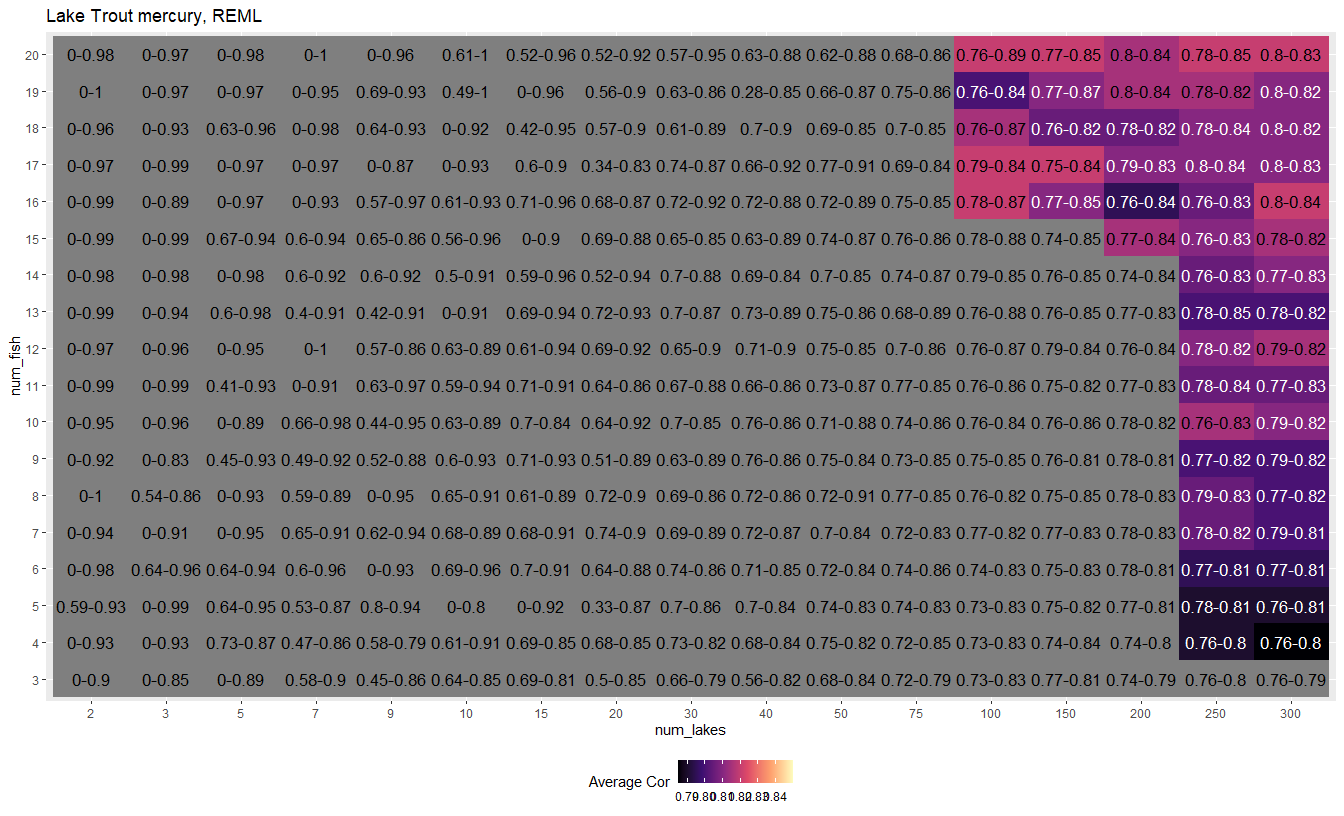


Figure 13: Range of Pearson correlation scores for simulated samplings of Lake Trout [Hg] REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish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

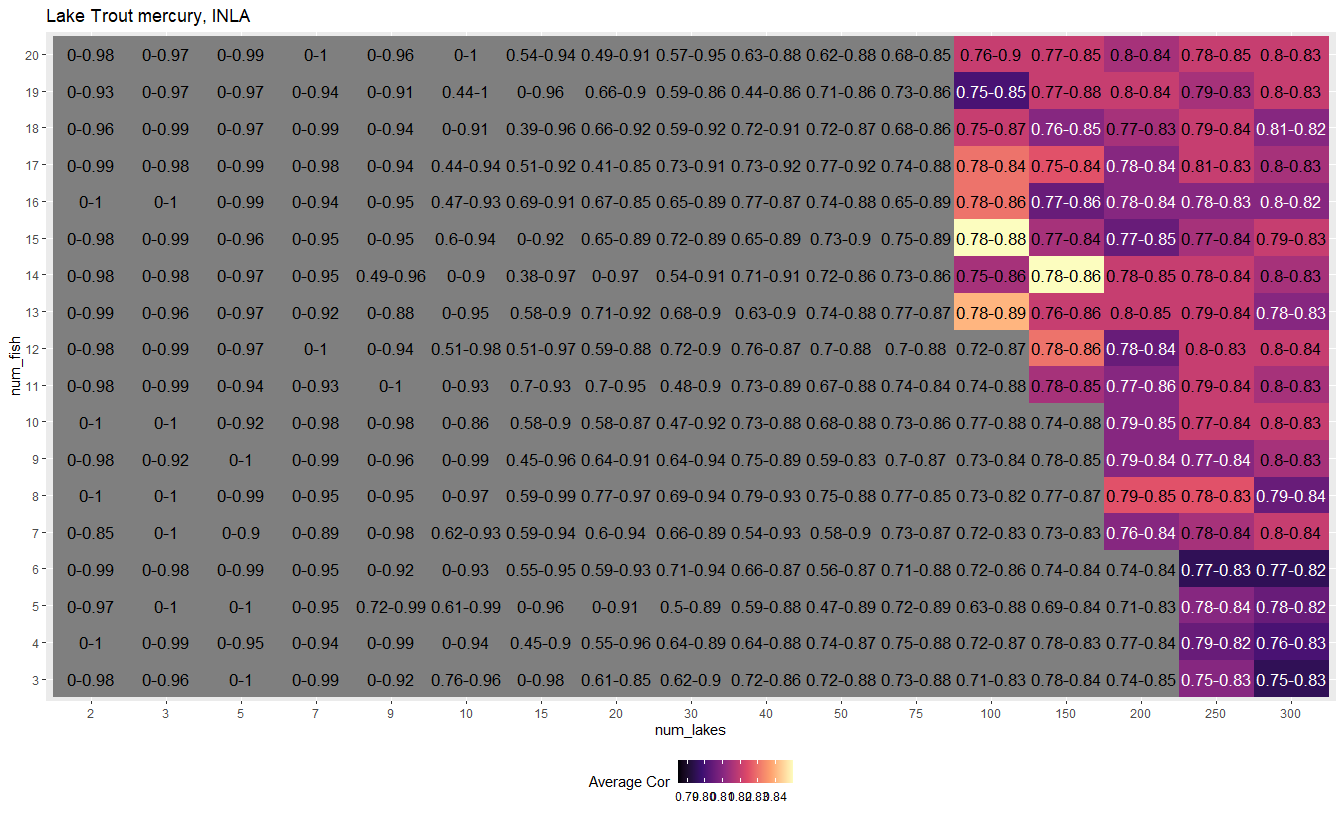


Figure 14: Range of Pearson correlation scores for simulated samplings of Lake Trout [Hg] INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish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

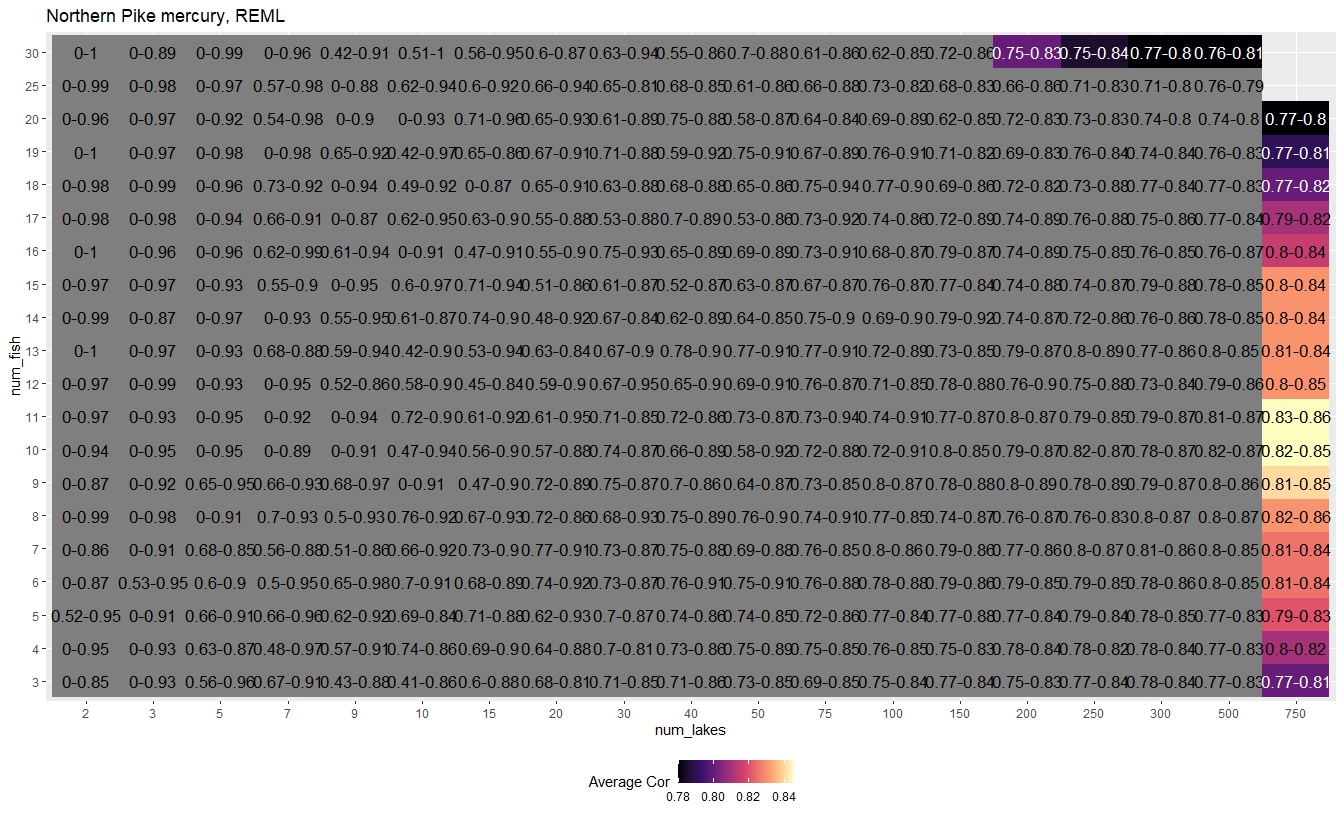


Figure 15: Range of Pearson correlation scores for simulated samplings of Northern Pike [Hg] REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish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

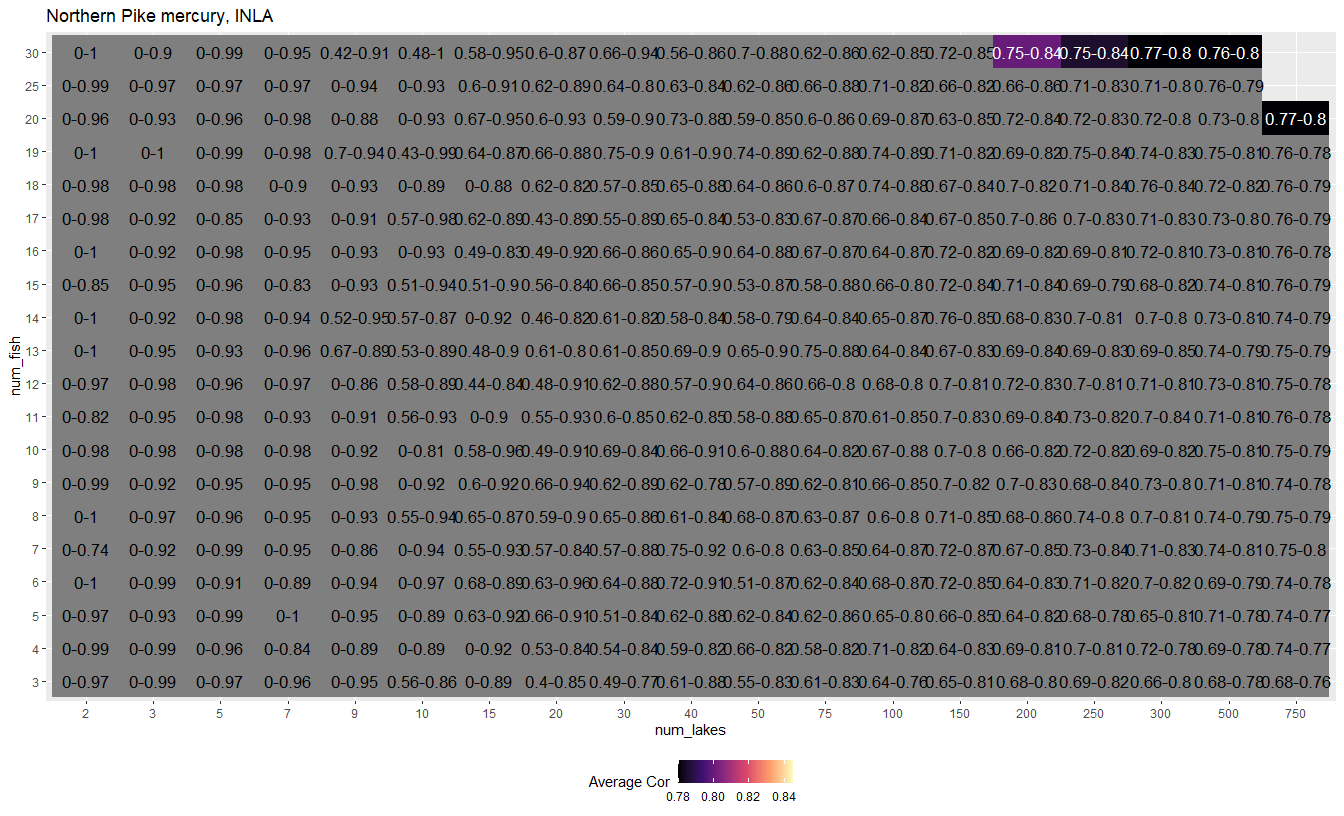


Figure 16: Range of Pearson correlation scores for simulated samplings of Northern Pike [Hg] INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish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

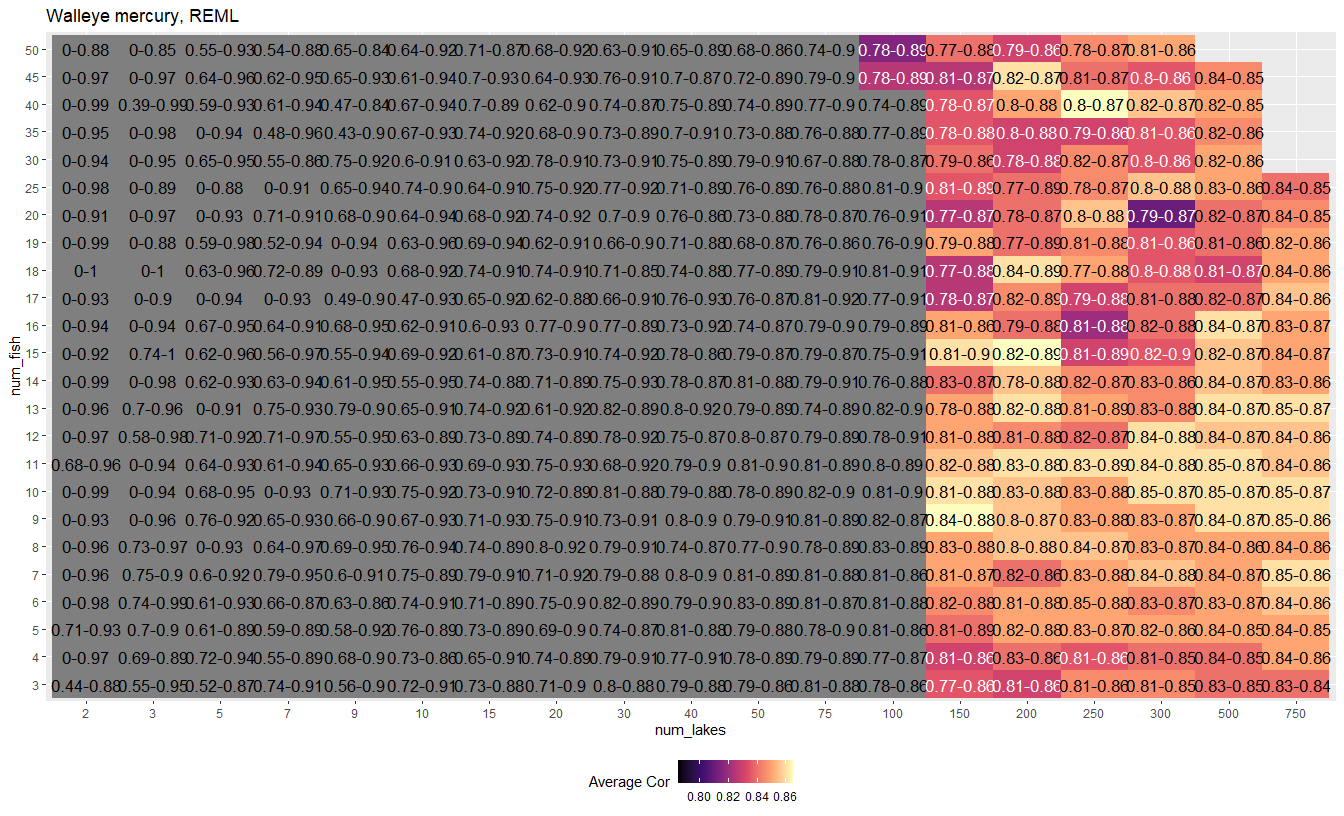


Figure 17: Range of Pearson correlation scores for simulated samplings of Walleye [Hg] REML predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or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

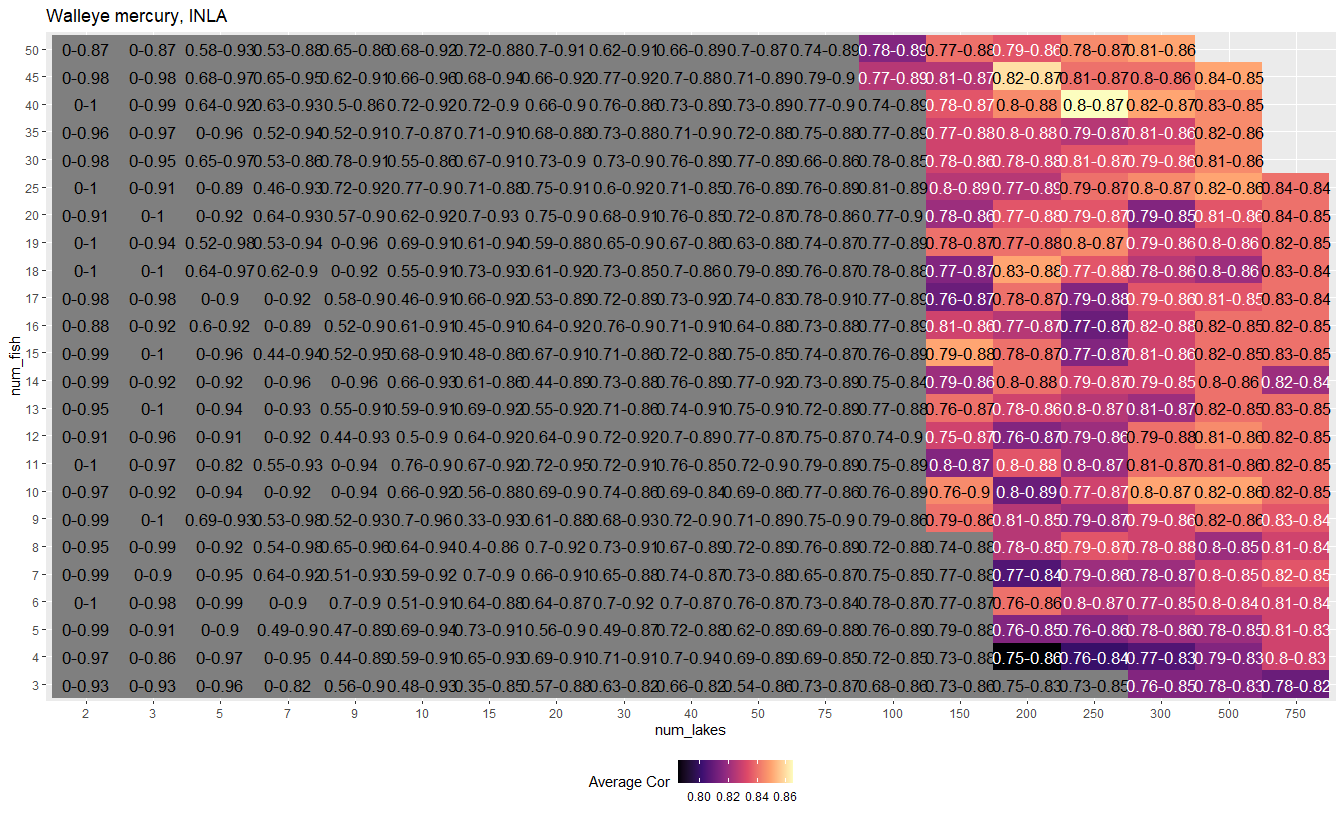


Figure 18: Range of Pearson correlation scores for simulated samplings of Walleye [Hg] INLA predictions from different lake and fish number combinations. Combinations where there is inadequate performance (min correlation < 0.75 and max correlation <.8) and where performance is unstable when fish or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

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.; Heerschap, M. J.; Keller, W. (Bill).; Southee, F. M.; 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.; Swanson, H. K.; Johnston, T.; Guildford, S. 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.; 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.; 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) 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.

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

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

(20) 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>.

(21) 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>.

(22) 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>.

(23) 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>.

(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) Lindgren, F.; Rue, H. [Bayesian Spatial Modelling with R-INLA](http://www.jstatsoft.org/v63/i19/). *Journal of Statistical Software* **2015**, *63* (19), 1–25.

(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>.