# **Assessing the potential responses of ten important fisheries species to a changing climate with machine learning and observational data across the province of Québec**

Marco A. Rodríguez+,1, Marine Lemaire+,1,2, Vincent Fugère 1, Marie-France Barrette3, Stéphanie Gagné3, Véronique Leclerc3, Olivier Morissette3,4, Rémy Pouliot3, Annick St-Pierre2, Katrine Turgeon5, Katherine Velghe2, Jean-Christophe Guay6, Beatrix E. Beisner2\*

Affiliations:

1. Département des sciences de l’environnement and Groupe de recherche interuniversitaire en limnologie (GRIL), Université du Québec à Trois-Rivières (UQTR), C.P. 500  
   Trois-Rivières, QC, G9A 5H7.
2. Département des sciences biologiques and Groupe de recherche interuniversitaire en limnologie (GRIL), Université du Québec à Montréal (UQAM), C.P. 8888 Succ. Centre-ville, Montréal, QC, Canada, H3C 3P8.
3. Ministère de l’Environnement, de la Lutte contre les Changements Climatiques, de la Faune et des Parcs, Direction principale de l’expertise sur la faune aquatique, 880, chemin Ste-Foy, Québec (Québec), G1X 4X4
4. Département des sciences fondamentales and Groupe de recherche interuniversitaire en limnologie (GRIL), Université du Québec à Chicoutimi (UQAC), 555, boulevard de l’Université, Chicoutimi, QC, G7H 2B1.
5. Département des sciences naturelles and Groupe de recherche interuniversitaire en limnologie (GRIL), Université du Québec en Outaouais (UQO), 283, boulevard Alexandre-Taché, C.P. 1250, Succursale Hull, Gatineau, QC, J8X 3X7.
6. Hydro-Québec, Direction Environnement, Expertise Milieu Naturel, 800 boul. de Maisonneuve Est 23e étage Montréal, Qc, H2L 4M8

+ Co-first authors

\* Corresponding author: Beatrix E. Beisner, email: beisner.beatrix@uqam.ca

ORCID:

Rodríguez: 0000-0003-0391-9170

Lemaire: 0000-0002-3122-3229

Fugère: 0000-0003-1877-5755

Morissette: 0000-0002-5037-9093

Turgeon: 0000-0002-2528-130X

Beisner: 0000-0001-6972-6887

# **ABSTRACT**

# Models are needed to predict changes in game fish abundances with respect to climatic factors undergoing change, but such models are often limited by data availability and the capacity of statistical methods to fit challenging ecological datasets. We use current methods in machine learning to describe the responses of ten fish species to climatic factors across Québec. We assembled a new province-wide, synthetic dataset of fish catches spanning almost 50 years and 6000 sites. Extreme Gradient Boosting (XGBoost) models revealed that climatic factors are more important predictors of trends in game fish catches than nuisance factors (sampling gear, time), lending support to collating other heterogeneous datasets for analyses. Mean annual temperature and precipitation were the most important drivers of species catches. Fish thermal preference guilds predicted primarily species responses to temperature, suggesting that warmer and wetter climates may not favour the same species. Despite the challenging nature of these datasets, XGBoost models provided excellent fit, predictive capacity, and interpretability, thereby illustrating that large, heterogeneous datasets can be used to inform freshwater fisheries management in a changing climate.

# **KEYWORDS**

# Extreme Gradient Boosting (XGBoost), SHAP values, synthetic dataset, freshwater fish, management

# **INTRODUCTION**

# How species will respond to global climate warming is a major focus of ecological research during these times of rapid change. In freshwaters, climate warming will be accompanied by changes in precipitation and hydrological regimes, including interactions with other anthropogenic stressors (e.g., land use change, fishing, invasive species), leading to cumulative effects that are challenging to predict (Schindler 2001). For inland fishes, these changes will impact individual growth, reproduction, survival, migration patterns and physiological adaptations. This should ultimately lead to change in species abundances and assemblages, and to shifts in geographical ranges and evolutionary trajectories (reviewed in Ficke et al. 2007; Myers et al. 2017; Barbarossa et al. 2021). These effects are known to vary based on species thermal traits, with coldwater species being generally more susceptible to climate warming than cool or warmwater species (Jones et al. 2013). Freshwater fishes will also respond to changing precipitation patterns (magnitude and timing), and thereby hydrological regimes of water bodies (Comte et al. 2013) but responses to these changes are less documented than are responses to warming (Myers et al. 2017).

# With respect to recreational fisheries, there are also economic and social motivations to develop better predictive models and an ability to anticipate population responses to global change (Lynch et al. 2016). With greater accessibility to empirical models, built upon species responses observed to date, the scientific and management communities will be better positioned to respond to potential future changes in recreational fisheries.

# Until recently, building such predictive models for fish communities has been limited both by access to sufficient data and the ability to build robust empirical models that can reliably predict responses to climatic factors. The data hurdle is one we are gradually overcoming as data are being increasingly shared in open-access repositories by a growing number of governmental and non-academic research agencies tasked with environmental monitoring for many decades. In our study region – the province of Québec, Canada – annual monitoring of recreational fisheries by the provincial government, follow-up studies of dam-building effects by the Hydro-Québec hydropower company, and impact assessments mandated by law and conducted by consulting firms are three important sources of species-level fish catch (individual fish counts) data that have not previously been combined and used to document species responses to climatic factors. With respect to climatic variables, efforts to distribute high quality global meteorological data, thereby enabling the development of predictive models, are being made via initiatives such as WorldClim (Fick and Hijmans 2017). Our research initiative (the FisHab Project) aimed at merging these three data sources into a common, large-scale fish catch database, to be used in conjunction with climate data to better understand the potential impacts of climate change across Québec, the largest province in Canada. Given the size of Québec (1 542 056 km2) and the range of latitudes covered (from 44°59'30"N to 62°34'55"N), the range of climatic variation is substantial (i.e., comprising humid continental, subarctic, and tundra climates in the Köppen classification), thereby providing an excellent study region over which to examine fish responses to climatic factors likely to change with global warming. In addition to climatic factors, we include in our analyses two landscape-level factors, elevation and habitat type, as potential predictors of fish catches. Elevation captures a broad variety of climatic, tectonic, geologic, and watershed features that can affect hydrologic behaviour (Jha and Diplas 2018), and in turn, influence fish responses. Elevational gradients can affect limiting nutrient concentrations (Beck et al. 2021), sediment dynamics (Sutfin and Wohl 2019), and stream habitat and water quality (Arocena et al. 2018). Elevation also reflects potential topographical barriers to dispersal and additional restrictions on species distributions, including historical constraints derived from biogeographical processes (Minns and Moore 1995; Warren et al. 2014). Finally, responses of streamflow to climate change may depend on, and require adjustments for, elevation (Tennant et al. 2015). Two habitat categories, lakes and rivers, were used in the analyses.

# While data limitations are being overcome, robust statistical approaches are also needed to analyse large, often heterogeneous synthetic datasets. Important advances in machine learning technology in recent years present new opportunities for ecological applications (Lucas 2020; Pichler and Hartig 2023). For example, the Extreme Gradient Boosting algorithm (XGBoost: Chen and Guestrin 2016) offers numerous advantages over less-current statistical methods that could make it a better algorithm to predict species abundances. These advantages include automatic inclusion of interactive and nonlinear effects of predictors, robustness to predictor collinearity, effective regularization to reduce overfitting, and fast performance when dealing with large datasets, including sparse datasets having a preponderance of zeros (Chen and Guestrin 2016). XGBoost’s versatility and outstanding predictive performance have led to its rapid adoption in many fields (Nielsen 2016), including recent applications in fish ecology (Muñoz-Mas et al. 2019; García-Gómez et al. 2023) and fisheries (Kaemingk 2020). Modelling species abundances in the form of count data often poses specific challenges arising from both marked over-representation of zero counts and overdispersion relative to a Poisson distribution (Fig. S1). Traditional statistical approaches to modelling such data include hurdle and zero-inflated models, which can incorporate both environmental variables and random effects in their linear predictor terms (Martin et al. 2005; Arab et al. 2008; Clark and Gelfand 2019). However, these models require explicit specification of all predictors of interest, including potential interaction terms, in both the occurrence and abundance components of the model. In contrast, the efficient handling of sparsity, automatic inclusion of predictor interactions, and unified treatment of occurrence and abundance by XGBoost may provide a simpler means to interpret environmental effects on species abundances.

# Here we employ XGBoost and our new synthetic database of fish catches across Québec to analyse the responses of ten game fish species to climatic and landscape factors, time, and various nuisance factors related to sampling (fishing gear, time of the year). All ten focal species (lake sturgeon (*Acipenser fulvescens*), lake whitefish (*Coregonus clupeaformis*), northern pike (*Esox lucius*), yellow perch (*Perca flavescens*), arctic char (*Salvelinus alpinus*), sauger (*Sander canadensis*), lake trout (*Salvelinus namaycush*), brook trout (*Salvelinus fontinalis*), Atlantic salmon (*Salmo salar*), and walleye (*Sander vitreus*) are important for recreational or subsistence fisheries in our study region. To test whether species thermal affinities influence their response to climatic factors, we divided species into thermal preference guilds (‘cold-water’, ‘cold/cool-water’ and ‘cool-water’ species) and analysed responses based on these (Coker et al. 2021). We additionally compared model-derived species responses to habitat type with species preferences for lentic and lotic habitats documented in the literature (Scott and Crossman 1973). Our models aim to: 1) test the descriptive and predictive capacities of XGBboost models fitted to fish catch data from province-wide surveys; 2) identify the most important climatic and landscape variables driving the catch of these game fish species, including factors related to temperature and precipitation; and 3) determine whether warmer or wetter conditions might affect these species in the coming decades, with species-specific responses determined by species thermal preferences. As an additional objective, we also document temporal trends in game fish catches in fish surveys conducted across Québec over the past 40 years.

# **MATERIALS AND METHODS**

### ***Sampling sites and fish species***

The sampling sites covered most of the southern regions of Québec (Fig. 1). Data were provided by both Hydro-Québec and what is now known as the Québec *Ministère de l’Environnement, de la Lutte contre les changements climatiques, de la Faune et des Parcs* (MELCCFP) but was formerly two separate ministries of environment (MELCC) and wildlife conservation and management (MFFP) (Fig. 1). These datasets include a heterogeneous mix of standardized and non-standardized fish surveys by government biologists and by environmental consulting firms sub-contracted by Hydro-Québec, collected between 1973 and 2021. Overall, 6498 unique sites (3087 sites in lakes, 3412 sites in rivers; reservoirs excluded), were included in the database. For each site, the data consisted of species counts (adult and juvenile life stages), location (latitude and longitude), sampling date, habitat type (lake or river) and fishing gear (three categories: electrofishing, gillnet, or seine).

The analysis focused on ten fish species important for recreational and subsistence fisheries in Québec (Table 1). In addition to their relevance for fisheries management, these ten species are easy to identify morphologically, which lessens a source of variance (taxonomic expertise of fisheries biologists and technicians) across the datasets that we grouped. We assigned species to thermal preference (preferenda) categories following Coker et al. (2001) (Table 1). Under this classification, northern pike, yellow perch, sauger, and walleye are considered ‘cool-water’ species, lake sturgeon are considered a ‘cold/cool-water’ species, and lake whitefish, arctic char, lake trout, brook trout, and Atlantic salmon are considered ‘cold-water’ species. We also assigned species to habitat preference categories ("lentic", "lotic", or "none") based on Scott and Crossman (1973) (Table 1). The distribution of catches for these ten species (Fig. S1) makes them challenging to analyse with traditional statistical methods used in ecology.

### ***Climate and elevation data***

Climate data (means over 1970 to 2000) and elevation were extracted for each site according to the site’s location from the WorldClim website (<https://www.worldclim.org/>), which provides global weather and climate data at high spatial resolution. All 19 ‘bioclimatic variables’ (O’Donnell and Ignizio 2012) provided by WorldClim were obtained from WorldClim raster files (10-min. resolution) using the function *extract* from the package *raster* (Hijmans and van Etten 2012). To facilitate interpretation, only five weakly-correlated bioclimatic variables were retained for analysis (Table 2). Two of these variables represented annual means for temperature and precipitation, whereas the remaining three represented intra-annual temporal variation: mean diurnal temperature range, annual temperature range, and precipitation seasonality (coefficient of variation of monthly total precipitation).

***Species catch models***

XGBoost (extreme gradient boosting) is a machine learning algorithm that relies on a sequential combination of simple decision trees (‘weak learners’), in which residual errors left by earlier trees are reduced by subsequent trees in the sequence, with improvements being guided by a gradient-descent algorithm that targets an objective function. The resulting ensemble of weak models yields a collectively strong model (Chen and Guestrin 2016).

Here, we built a model for each target species,i.e., ten models in all. For each model, the response variable was the catch (counts) of the focal species, and the predictors were site elevation, habitat type (lake or river), fishing gear (gill net, seine, or electrofishing), month, year, and the five bioclimatic variables described in the previous section (Table 2). Habitat type was coded as a binary variable assuming a value of 1 for lakes and 2 for rivers. Fishing gear was coded by means of two binary (0 or 1) dummy variables: gillnet, assuming a value of 1 for gillnets and 0 otherwise, and seine, assuming a value of 1 for seines, and 0 otherwise, with electrofishing as the baseline or reference category (Fahrmeir et al. 2013). Including fishing gear and time (year, season) as predictors allowed us to compare the importance of climatic and landscape variables relative to that of methodological variables that can be considered ‘nuisance’ variables but are known to influence catch.

We used a Bayesian optimization procedure to tune the XGBoost model hyperparameters (Garnett 2023). Specifically, we used the algorithm proposed by Bischl et al. (2017) as implemented in the R package ‘mlrMBO’ to optimize the following hyperparameters (intervals sampled shown in parentheses): learning rate (eta; 0.005 - 0.4), minimum loss reduction required to make a split (gamma; 0 - 50), maximum tree depth (max\_depth; 3 –10), minimum sum of weights of observations required in a child node (min\_child\_weight; 1 -100), subsample ratio of the training instances (subsample; 0.4 - 1), and subsample ratio of columns when constructing each tree (colsample\_bytree; 0.4 - 1). We used expected improvement (Garnett 2023) as the utility function in the optimization procedure. For each species, we specified an initial design by placing 50 points in the parameter hyperspace following a space-filling Latin hypercube design and ran the optimization algorithm for 100 iterations (or “epochs”), which was sufficient to ensure that the utility function stabilised, as determined from graphical inspection of declines in utility. A pseudo-*R*² based on the Poisson likelihood (Mittlböck and Waldhör 2000) allowed us to assess the fit of our final species models. We additionally used a pseudo-*R*² calculated from omitted samples in a cross-validation procedure to assess the out-of-sample predictive performance of the final models.

The fitting procedure we used for each species model can be summarized as follows: 1) provide initial estimates for the hyperparameters and 2) fit the model under that set of hyperparameters, using the cross-validation procedure implemented in function ‘xgb.cv’ from the ‘xgboost’ R package (Chen et al. 2023). We carried out cross-validation using ten random partitions of the dataset to fit individual models using a Poisson negative log-likelihood as the evaluation metric. Then: 3) evaluate expected improvement under the current set of hyperparameters and propose a new set of hyperparameters; 4) iterate through steps 2 and 3 until the utility function stabilizes, yielding an optimal set of hyperparameters; 5) fit the final model using function ‘xgboost’ and the optimal set of hyperparameters, and quantify model fit using the pseudo-*R*² value; 6) refit the model using function ’xgb.cv' and the optimal set of hyperparameters to obtain a cross-validated pseudo-*R*² value. The pseudo-*R*² value of the final model is an appropriate measure of fit for count data, whereas the cross-validated pseudo-*R*² provides a better measure of the final model's predictive capacity. Finally, 7) we used SHAP (SHapley Additive exPlanations) values (Lundberg and Lee 2017) to interpret the relationships between variables and model predictions of species catch using the R package ‘SHAPforxgboost’ (Liu and Just 2020). SHAP values measure the contribution of each individual predictor to the model's predictions; they are derived from game theory principles and have a solid theoretical background. In contrast to other commonly used measures of predictor importance, such as “gain” (average improvement in loss brought by a predictor), SHAP values uniquely satisfy some desirable properties. These include local accuracy (accurate predictor contributions are available for individual observations), missingness (predictors that are missing or irrelevant for prediction have zero values), consistency (if model changes so that a predictor’s contribution increases or remains the same, the predictor’s SHAP value also increases or remains the same), and additivity (predictor contributions are calculated independently and then summed to obtain the prediction) (Lundberg and Lee 2017). This allowed us to calculate the importance of the potential driving variables in each of the species models and explore the direction and variability of responses (represented on a logarithmic scale as positive or negative deviations from the mean predicted catch; see below) to each of the driving variables. For any given observation, the SHAP value of a predictor indicates the sign and magnitude of contribution of the predictor to the value predicted for that observation. The relationships between SHAP values and predictors were visualised in dependence plots with loess smoothers (‘loess.smooth’ function in R) for the continuous predictors, and with boxplots of SHAP values and predictor categories for a single discrete predictor, habitat type. The dependence plots provide a flexible representation of potentially complex relationships between individual predictors and the response variable, and the vertical scatter around the main trends in these plots provides insight into potential interactions between the focal predictor and other predictors (Lundberg et al. 2020). We used this feature of dependence plots to determine how observed species responses to climatic and landscape predictors differed among fishing gears. Specifically, we used dependence plots identifying species responses by gear type to check for null (overlapping trends for different gears), additive (parallel non-overlapping trends), and interactive (non-parallel trends) effects involving gear type.

For each species, we also calculated two metrics of habitat preference and compared them to the habitat preference categories derived from the literature. The first metric is model-based and is calculated as the difference in mean SHAP values of lakes and rivers, whereas the second metric is obtained as the natural logarithm of the ratio of mean catches of lakes and rivers. Both metrics quantify preference for lakes over rivers and should therefore decline monotonically across habitat preference categories ordered along the sequence "lentic-none-lotic".

The connection between SHAP values and model predictions can be made more explicit. Assuming that each observation has a set of *M* associated predictors , the local accuracy and additivity properties mean that:

where is the model prediction for observation , a constant "bias" term, is the average prediction over the data set, and is the SHAP value associated with observation and predictor . The SHAP values explain the difference between the predicted value and the overall average prediction . Each SHAP value therefore represents the (positive or negative) contribution of a single predictor to the predicted value for an individual observation relative to the overall average. In our XGBoost models the predicted are Poisson means modelled on the logarithmic scale, so that . This means that the predicted values on the natural scale, , are directly linked to the bias term and SHAP values through the relation . The importance of individual predictor *j*, *I*j, was quantified as the mean of the absolute SHAP values for the predictor across the data:

To compare the performance of XGBoost with a more conventional approach to species distribution modelling, we also analyzed the same dataset using a zero-inflated negative binomial (ZINB) regression model; an approach commonly used to deal with excess zeros and overdispersion in count data (Lewin et al. 2010; Stoklosa et al. 2022). The 11 predictors listed in Table 2 were included as fixed effects in the abundance component of the model, with the exception of "year" and "month", which were treated as random effects. The occurrence component of the model only included the three most important climatic variables identified by the SHAP values, annual mean temperature, annual precipitation, and precipitation seasonality, as attempts to include additional fixed or random predictors in this component led to non-convergence for several species. The ZINB model was fitted using R package ‘glmmTMB’ (Brooks et al. 2017). We compared the fit of XGBoost and ZINB models by means of four metrics that measure the consistency between observed and predicted catch: mean absolute error and root-mean-square error (both scaled by dividing the error by the mean observed catch), the coefficient of variation (squared Pearson correlation), and the Spearman rank correlation *ρ* (Zhang et al. 2020, Waldock et al. 2022).

# **RESULTS**

The XGBoost models fitted these challenging data surprisingly well (Table 1; Fig. S2). Ranking predictors using SHAP values showed that climatic variables were the most important explanatory variables for most species (Fig. 2). Responding most strongly to temperature variables were cold-water species: Arctic char, lake whitefish and Atlantic salmon (Fig. 2 top row), but only one cool-water species – sauger (Fig. 2 bottom row). Moreover, annual mean temperature was the most important climatic variable, although a secondary one overall, for cold-water lake trout and cool-water yellow perch. Responding most strongly to precipitation variables were cool-water species such as the Northern pike and walleye, but also the cold-water brook trout. For the cold/cool-water lake sturgeon, precipitation variables were a secondary explanatory variable, followed by mean annual temperature (Fig. 2). For lake sturgeon, elevation emerged as the first explanatory variable. Habitat type was most important for yellow perch and lake trout (followed by elevation in the latter case). Sampling date and fishing gear types were generally less important predictors than climatic variables, except for lake whitefish (Fig. 2).

We further summarized the results of variable importance from Figure 2 using boxplots of variable importance across the ten species models (Fig. 3). The ordering by decreasing mean importance (rank) demonstrates that the climatic variables were generally the most important across species, with annual means for temperature and precipitation ranking as the two most important drivers. Moreover, annual mean temperature showed the least variation among species of the climatic variables (Fig. 3). Again, the lower ranking of importance for fishing gear and sampling date are clearly visible.

We also used dependence plots (Figs. 4 - 8) to examine the relationship between modelled catches (SHAP values) and each of the five top-ranked drivers when pooling all models: annual mean temperature, annual precipitation, the seasonality of precipitation, elevation, and habitat (Fig. 3). In the dependence plots, the range of SHAP values along the Y axis for a given species and predictor is related to the importance value of the predictor as shown in Fig. 2. Known cold water species generally showed declines with increasing temperature (Fig. 4 – top row), while known cool-water species generally increased (Fig. 4 – bottom row). With our modelling approach it is also possible to estimate temperature values at which distinct species declines or increases are more likely to occur. For example, lake whitefish begins declining at much cooler mean annual temperatures (~ -2.5°C) than does brook trout (~ +2.5°C) and the other three cold-water species, which decline from the lowest observed temperatures (Fig. 4 – top row). For cool-water species, most increased across the temperature range observed, but especially at temperatures above 0 °C.

For annual mean precipitation, we again examined responses for known cold- (Fig. 5 - top row) and cool-water species (Fig. 5 – bottom row). While annual precipitation was identified as a top explanatory variable (Fig. 3), species-specific responses were less linear than for annual temperature, and less predictable from species thermal traits (Fig. 5). In general, very high precipitation was associated with lower abundances in all species except brook trout. For both cold- and cool-water species, increasing precipitation generally led to a bell-shaped species response, with catches peaking at intermediate levels of annual mean precipitation (800 - 1200 mm/yr).

Increasing seasonality of precipitation led to catch declines in four species (lake trout, brook trout, Atlantic salmon, and yellow perch), to generally higher catches in three species (lake sturgeon, walleye, and northern pike), and to more strongly non-linear patterns in the remaining three species (Arctic char, lake whitefish, and sauger; Fig. 6). As seen for annual precipitation, the response to precipitation seasonality did not depend on temperature preferences. Elevation was an important variable identified for cold-water lake trout, brook trout and for lake sturgeon, sauger, and walleye (Fig. 2). Generally, the cool-water species showed declines at high elevation (Fig. 7 – bottom row) while cold-water lake trout and lake whitefish increased with elevation, up to a point, and brook trout showed a nearly linear increase with elevation (Fig. 7 – top row). Responses to habitat type suggested greater use of lake habitat for 8 of the 10 species; only Atlantic salmon and lake sturgeon showed positive responses to river habitat, as shown by the SHAP dependence plots (Fig. 8) and the two metrics of habitat preference (Fig. 9; top). Although the two metrics were well correlated (Spearman *ρ* = 0.94), the habitat preference metric based on SHAP values was less variable than the metric based on the log-ratio of mean catches (Fig. 9; top). On average, the SHAP-based metric showed less variation within categories than the catch-based metric (Fig. 9; bottom). The SHAP-based metric also indicated a marked preference for lakes by lake trout and yellow perch (Fig. 9; top), the two species for which habitat type was the most important predictor (Fig. 2).

Finally, we also were able to look at long-term trends in catches, with all species showing generally positive trends between 1973 and 2021 (Fig. S3). Year was among the top explanatory variables only for Atlantic salmon (Fig. 2), and was overall one of the least influential predictors in the model (Fig. 3). The overall increases observed for most species over time did not appear to be related to temporal variation in the latitude of sampling locations, which remained stationary over most of the sampling period (Fig. S4).

Dependence plots identifying species responses by gear type yielded little indication that species responses were driven by sampling artifacts related to gear characteristics (Figs. S5 - S10). Gear effects appeared to be mostly null (overlapping parallel trends) or additive (non-overlapping parallel trends), with any interactions (non-parallel trends) modifying the magnitude but not the overall direction of the trend (e.g., lake sturgeon and elevation; lake whitefish and year) (Figs. S5 - S10).

The ZINB model performed poorly with these challenging data (Table S1). For all species, the consistency between predicted and observed values was considerably better for the XGBoost model.

# 

# **DISCUSSION**

Our study assembled an extensive fish catch dataset, covering the entire province of Québec, and analyzed it for the first time using XGBoost, a powerful new machine learning approach. The models generated by XGBoost had good 1) descriptive power, as judged by the pseudo-*R*² and graphical comparisons of observed and fitted values, 2) predictive capacity, as gauged by the pseudo-*R*², and 3) interpretability, via the SHAP values and their associated dependence plots. Interestingly, variables related to sampling (fishing gear, sampling month) had overall little effect on the responses observed in the study region, which supports the idea of assembling large synthetic databases encompassing multiple fishing gears and surveys for bioclimatic analyses such as the one we have presented. Our comparison of the XGBoost and ZINB models is clearly intended to be illustrative rather than comprehensive. However, the comparison highlights some potential advantages of XGBoost as a broadly applicable tool that can help simplify or entirely sidestep some of the decisions required by more conventional approaches. Examples include the choice of predictors (fixed or random, additive or interactive, transformed or raw) and model structure (e.g., zero-inflated, hurdle, or neither), both of which may be species-specific and substantially influence the effort required to achieve model convergence and proper fit.

Overall, climatic predictors emerged as the most important in explaining the catch of most game fish species examined. Of the climatic predictors, annual mean temperature and annual precipitation were the most important, followed by the non-climatic predictor elevation. These results do not mean that proximate local variables (e.g. nutrients, turbidity, current flow, substratum type, water temperature, dissolved oxygen) would not also play an important role in defining population productivity and abundance (e.g. see Knouft and Anthony 2016; Jarvis et al. 2020), but they do point to ways in which different game fish may be more or less sensitive to temperature and precipitation shifts anticipated with climate change across a large landscape, covering a wide range of latitudes (spanning 17.1 ° and 15.5 °C in annual mean temperature in our study). The performance of our XGBoost models show that at this spatial scale, climatic data that are readily available online in open access are sufficient to better understand and predict species abundance variations. Previous work has shown that species distribution models using abundance data, even at a coarse scale, can lead to better predictive capacity and identification of relative habitat suitability than models using only occurrence data (Howard et al. 2014). Similarly, macroclimatic variables can be better predictors of ecological processes than smaller-scale predictors such as local climate (Hallett et al. 2004), and have been shown to be useful proxies for local environmental predictors of abundance in freshwater fishes (Frederico et al. 2014; Knouft and Anthony 2016).

Generally, our analyses support the classification of cold and cool-water species based on their responses to mean annual temperature. Species responses to this variable lined up with their thermal preferences reported in the literature: coldwater species declined with increasing temperature, whereas cool-water species increased with increasing temperature. Interestingly, among the cool-water species only yellow perch and sauger responded to temperature as a primary climatic driver, with both species showing almost linear increases in catch with warmer temperatures.

Rather than primary responses to temperature, cool-water species catches tended to respond more often to precipitation as a primary climatic driver, as observed for northern pike, lake sturgeon, and walleye. Generally, the catch of all cool-water species responded favourably to increasing precipitation to a certain maximum (800-1000 mm/yr) and then tended to decline; conditions that were also associated with greater seasonality in precipitation across the landscape. Indeed, in the study region, mean annual precipitation is negatively correlated with seasonality (Pearson *r* = ‑0.85). For cold-water species that responded (all except Arctic char), the maximum catches were generally observed at greater annual precipitation values (1000-1200 mm/yr), also associated with lower seasonality. For these cold-water species, lower seasonality would mean more reliably wet conditions throughout the year and less chance of water-level reductions or even drought.

For the more southernly species, one might expect greater precipitation values to be more destructive given that more precipitation will fall as rain (as opposed to snow in the sites where cold-water species occur). Thus, the declines observed at high annual precipitation levels could be associated with negative effects of flooding occurring throughout the year. Flooding in streams can perturb fish habitat directly (Suttle et al. 2004), and have strong impacts on fish abundances (Carline and McCullough 2003; George et al. 2015). More generally, changes in precipitation that modify the river flow regime can influence the severity of summer droughts, reproductive success, and seasonality of movements, all of which impinge on the local abundances of stream fish (Magalhaes et al. 2007; Rayner et al. 2009; Warren et al. 2015; Kanno et al. 2016). In lakes, large inputs with flooding events or runoff events can also reduce fish habitat quality by bringing in sediments that can darken waters or interfere with gill functioning (Donohue and Garcia-Molinos 2009), and increased nutrient loading can lead to shifts in ﬁsh community structure toward small and abundant plankti-benthivorous ﬁsh (Jeppesen et al. 2009). Where greater snowpack occurs, these flooding events will be more restricted to a single timepoint or may be moderated by melting rates, with less of an overall negative impact on fish habitat and therefore their abundances (Ficke et al. 2007).

In addition to climatic variables, elevation, and habitat type (river vs. lake) emerged as important for a few species. As would be expected, the cold-water brook trout responded positively to elevation – likely benefitting from both colder temperatures, but also lower-order streams at higher elevation (Weigel and Sorensen 2001). Similarly, it is not unexpected that cool-water species would be favoured at lower elevations given that will be exposed to warmer temperatures in this way. Lake trout abundances also responded to elevation (after habitat type), but in a unimodal way, by first increasing and then declining at elevations >500 m. Lake trout populations will be favoured at colder temperatures, but across Québec these often occur also in the hypolimnia of deep lakes that occur in lowland areas of the province. In all cases, elevation can influence a variety of other habitat characteristics important for fish populations including connectivity and accessibility (Carvajal-Quintero et al. 2019) as well as factors associated with landscape position in the hydrographic network, such as waterbody surface area (Martin and Soranno 2006), productivity (Davis et al. 2014), and allochthonous inputs (Kortelainen 1993; Martin and Soranno 2006) and food web structure (Reid et al. 2011) among others.

Together these results point to likely population declines in several important game fish with increasing temperatures associated with climate change, especially among coldwater species. Increases in annual mean temperature in Quebec over the study period are small relative to those predicted over the next 80 years under various simulated scenarios (Ouranos 2015, Fig. 1.6). The most northern-occurring species, Arctic char, is likely to experience the greatest population declines followed by lake trout. Atlantic salmon, brook trout, and lake whitefish will also be somewhat sensitive to temperature increase. On the other hand, a double jeopardy is faced by brook trout, which may suffer stronger population declines if droughts occur in conjunction with warming. These species are likely vulnerable to replacement by expanding populations of yellow perch, sauger, and walleye with warming temperatures and where species dispersal can occur via connected habitats, in agreement with numerous studies foreseeing range shifts and replacement of cool/cold water species with warmwater species (Shuter and Post 1990; Chu et al. 2005; Reist et al. 2006, Alofs et al. 2014; Poesch et al. 2016; Van Zuiden et al. 2016). Changing precipitation, however, has received less attention than warming in previous studies of fish responses to climate change (but see, e.g., Ficke et al. 2007; Knouft and Anthony 2016; Myers et al. 2017). Walleye will require moderate levels of precipitation to thrive. Lake sturgeon and northern pike also appear to be sensitive to very high levels of precipitation, and thus their expected positive responses to temperature could be modulated by the shift in mean precipitation regimes in response to climate change across the province of Québec.

The habitat preference metric based on SHAP values showed less variability than the metric based on the log-ratio of mean catches and seemed to reliably capture patterns of habitat use in our study region. Only one species showed a discrepancy between the SHAP-based metric and preferences derived from the literature: walleye, classified as "lentic" had an intermediate metric value similar to those of species classified as "none". Metric values for the three other species classified as having defined habitat preferences were as expected, positive for "lentic" species and negative for the single "lotic" species. Despite the small number of species in our comparison, the means of the SHAP-based metric for the "lentic", "none", and "lotic" categories were as expected positive, near-zero, and negative, respectively. Synoptic studies of fish ecology incorporating both lakes and rivers are rare (but see Randall et al. 1995; Irz et al. 2006; Wurtsbaugh et al. 2015). Although it is inherently difficult to compare preferences across habitats that differ in physical structure, spatial extent, and connectivity, such as lakes and rivers, the use of SHAP-based metrics resulting from the analysis of large synthetic databases appears to be a useful tool for quantifying habitat preference in freshwater fishes.

The generally positive temporal trends observed for most species result from spatial aggregation over the entire province and small temperature increases relative to those expected under future climate change scenarios (Ouranos 2015), which makes the trends difficult to interpret in terms of changes in climate over the study period. Changes in temperature and precipitation within Quebec, both past and predicted, show strong variability in space and time (Guay et al. 2015). This means that a thorough analysis of the effect of climate change on temporal trends in fish catches will likely require the use of ecoregions or a fully spatio-temporal model that incorporates temporally varying predictors.

Our study supports the use of novel machine learning techniques such as extreme gradient boosting (XGBoost) in combination with SHAP values to enhance interpretability. This approach enabled the identification of general groups of potential explanatory variables despite large differences in fishing techniques and approaches by different agencies and studies within agencies, noise in the data to which the analyses were relatively robust. We recommend continued work along these lines, especially where in-lake environmental data may also be present, to improve our ability to predict critical factors for fish populations and communities across biogeographical scales where a variety of disparate databases could be combined.

**COMPETING INTERESTS STATEMENT**

The authors declare no competing interests.

**AUTHOR CONTRIBUTION STATEMENT**

The project was conceptualized by MAR, VF, and BEB. Data was collected, collated and curated by ASP, KV, SG, JCG, and VL. Statistical analyses were done by MAR, VF, and ML. Writing of the original draft was done by ML, BEB, MAR, and VF with reviewing and editing by all authors. Project administration was carried out by BEB, KV, and ASP while all other authors participated in regular project meetings.

**FUNDING STATEMENT**

This research component of the FisHab project was funded by the Department of Fisheries and Oceans Canada and by the Québec Fonds de recherche – nature et technologie (FRQNT) via a grant to the Groupe de recherche interuniversitaire en limnologie (GRIL).

**DATA AVAILABILITY STATEMENT**

For the moment the data are proprietary, belonging to Hydro-Québec and the Québec Ministère de l’Environnement, de la Lutte contre les Changements Climatiques, de la Faune et des Parcs and can be shared upon request to those co-authors (JCG for Hydro-Québec at <guay.jean-christophe@hydroquebec.com> and SG for MELCCFP at <Stephanie.Gagne@mffp.gouv.qc.ca>). Metadata associated with this study are available on Borealis at <https://doi.org/10.5683/SP3/AHKTZJ>.

**REFERENCES**

Alofs, K.M., Jackson, D.A. and Lester, N.P. 2014. Ontario freshwater fishes demonstrate differing range‐boundary shifts in a warming climate. Diversity and Distributions, **20**(2):123-136.

Arab, A., Wildhaber, M.L., Wikle, C.K. and Gentry, C.N. 2008. Zero-inflated modeling of fish catch per unit area resulting from multiple gears: application to channel catfish and shovelnose sturgeon in the Missouri River. North American Journal of Fisheries Management, **28**(4), 1044-1058.

Arocena, R., Chalar, G. and Pacheco, J.P. 2018. Agriculture and elevation are the main factors for Pampasic stream habitat and water quality. Environmental Monitoring and Assessment, **190**:1-22.

Barbarossa, V., Bosmans, J., Wanders, N., King, H., Bierkens, M.F., Huijbregts, M.A. and Schipper, A.M. 2021. Threats of global warming to the world’s freshwater fishes. Nature Communications, **12**(1):1701.

Beck, W.S., Rugenski, A.T. and Poff, N.L. 2021. Limiting nutrients drive mountain stream ecosystem processes along an elevation gradient. Freshwater Science, **40**(2):368-381.

Bischl B., Richter J., Bossek J., Horn D., Thomas J. andLang M. 2017. mlrMBO: A modular framework for model-based optimization of expensive black-box functions. <https://arxiv.org/abs/1703.03373>.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H.J., Mächler, M. and Bolker, B. M. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal, **9**(2), 378-400.

Carline, R.F. and McCullough, B.J. 2003. Effects of floods on brook trout populations in the Monongahela National Forest, West Virginia. Transactions of the American Fisheries Society, **132**(5):1014-1020.

Carvajal-Quintero, J., Villalobos, F., Oberdorff, T., Grenouillet, G., Brosse, S., Hugueny, B., Jézéquel, C. and Tedesco, P.A. 2019. Drainage network position and historical connectivity explain global patterns in freshwater fishes’ range size. Proceedings of the National Academy of Sciences, **116**(27), pp.13434-13439.

Chen, T. and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). <https://doi.org/10.1145/2939672.2939785>

Chen T., He T., Benesty M., Khotilovich V., Tang Y., Cho H., Chen K., Mitchell R., Cano I., Zhou T., Li M., Xie J., Lin M., Geng Y., Li Y., Yuan J. 2023. xgboost: Extreme Gradient Boosting. R package version 1.7.5.1, https://CRAN.R-project.org/package=xgboost.

Chu, C., Mandrak, N.E. and Minns, C.K. 2005. Potential impacts of climate change on the distributions of several common and rare freshwater fishes in Canada. Diversity and Distributions, **11**(4):299-310.

Clark, J.S. and Gelfand, A.E. 2019. Accommodating so many zeros: univariate and multivariate data. In: Handbook of Environmental and Ecological Statistics, pp. 211-239. CRC Press.

Coker, G.A, Portt, C.B. and Minns, C.K. 2001. Morphological and ecological characteristics of Canadian freshwater fishes. Canadian Manuscript Report of Fisheries and Aquatic Sciences 2554: iv+89 pp.

Comte, L., Buisson, L., Daufresne, M., and Grenouillet, G. 2013. Climate‐induced changes in the distribution of freshwater fish: observed and predicted trends. Freshwater Biology, **58**(4), 625-639.

Davis, B., Mattone, C. and Sheaves, M. 2014. Bottom-up control regulates patterns of fish connectivity and assemblage structure in coastal wetlands. Marine Ecology Progress Series, **500**, 175-186.

Donohue, I. and Garcia Molinos, J. 2009. Impacts of increased sediment loads on the ecology of lakes. Biological Reviews, **84**(4), 517-531.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. 2013. Regression Models. Springer.

Frederico, R.G., De Marco Jr, P. and Zuanon, J. 2014. Evaluating the use of macroscale variables as proxies for local aquatic variables and to model stream fish distributions. Freshwater Biology, **59**(11):2303-2314.

Fick, S.E. and Hijmans, R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology **37**(12), 4302-4315

Ficke, A.D., Myrick, C.A., and Hansen, L.J. 2007. Potential impacts of global climate change on freshwater fisheries. Reviews in Fish Biology and Fisheries, **17**, 581-613.

García-Gómez, R.E., Aceves-Medina, G., Villalobos, H., Rosenberg, S.P.A.J. and Durazo, R. 2023. Predictive performance from abundance distribution models of *Vinciguerria lucetia* larvae in the southern portion of the California current system using XGBOOST. Deep Sea Research Part II: Topical Studies in Oceanography, **212**, p.105336.

Garnett, R. 2023. Bayesian Optimization. Cambridge University Press.

George, S.D., Baldigo, B.P., Smith, A.J. and Robinson, G.R. 2015. Effects of extreme floods on trout populations and fish communities in a Catskill Mountain river. Freshwater Biology, **60**(12):2511-2522.

Guay, C., Minville, M. and Braun, M. 2015. A global portrait of hydrological changes at the 2050 horizon for the province of Québec. Canadian Water Resources Journal/Revue canadienne des ressources hydriques, **40**(3):285-302.

Hallett, T.B., Coulson, T., Pilkington, J.G., Clutton-Brock, T.H., Pemberton, J.M. and Grenfell, B.T. 2004. Why large-scale climate indices seem to predict ecological processes better than local weather. Nature, **430**(6995), 71-75.

Hijmans, R.J. and van Etten J. 2012. raster: Geographic analysis and modeling with raster data. R package version 2.0-12.  [http://CRAN.R-project.org/package=raster](http://cran.r-project.org/package=raster)

Howard, C., Stephens, P.A. Pearce-Higgins, J.E., Gregory, R.D. and Willis, S.G. 2014. Improving species distribution models: the value of data on abundance. Methods in Ecology and Evolution, **5**(6), 506-513.

Irz, P., Odion, M., Argillier, C. and Pont, D. 2006. Comparison between the fish communities of lakes, reservoirs and rivers: can natural systems help define the ecological potential of reservoirs?. Aquatic Sciences, 68(1):109-116.

Jarvis, L.A., McMeans, B.C., Giacomini, H.C. and Chu, C. 2020. Species-specific preferences drive the differential effects of lake factors on fish production. Canadian Journal of Fisheries and Aquatic Sciences, **77**(10), 1625-1637.

Jeppesen E, Kronvang B, Meerhoff M, Søndergaard M, Hansen KM, Andersen HE, Lauridsen TL, Liboriussen L, Beklioglu M, Özen A, Olesen JE. 2009. Climate change effects on runoff, catchment phosphorus loading and lake ecological state, and potential adaptations. Journal of Environmental Quality, **38**(5):1930-41.

Jha, R. and Diplas, P. 2018. Elevation: a consistent and physically-based framework for classifying streams. Journal of Hydraulic Research, **56**(3):299-312.

Jones, R., Travers, C., Rodgers, C., Lazar, B., English, E., Lipton, J., ... and Martinich, J. 2013. Climate change impacts on freshwater recreational fishing in the United States. Mitigation and Adaptation Strategies for Global Change, **18**(2013),731-758.

Kaemingk, M.A., Hurley, K.L., Chizinski, C.J. and Pope, K.L. 2020. Harvest–release decisions in recreational fisheries. Canadian Journal of Fisheries and Aquatic Sciences, **77**, 194-201.

Kanno, Y., Pregler, K.C., Hitt, N.P., Letcher, B.H., Hocking, D.J. and Wofford, J.E. 2016. Seasonal temperature and precipitation regulate brook trout young‐of‐the‐year abundance and population dynamics. Freshwater Biology, **61**(1):88-99.

Knouft, J.H. and Anthony, M.M. 2016. Climate and local abundance in freshwater fishes. Royal Society Open Science, **3**(6):160093.

Kortelainen, P. 1993. Content of total organic carbon in Finnish lakes and its relationship to catchment characteristics. Canadian Journal of Fisheries and Aquatic Sciences, **50**(7), 1477-1483.

Lewin, W.C., Freyhof, J., Huckstorf, V., Mehner, T. and Wolter, C., 2010. When no catches matter: Coping with zeros in environmental assessments. Ecological Indicators, **10**(3):572-583.

Liu, Y. and Just, A. 2020. SHAPforxgboost: SHAP Plots for 'XGBoost'. R package version 0.1.0. https://github.com/liuyanguu/SHAPforxgboost/

Lucas, T.C. 2020. A translucent box: interpretable machine learning in ecology. Ecological Monographs, **90**, e01422.

Lundberg, S.M. and Lee, S.I. 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, **30**(2017).

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I. 2020. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, **2**(1):56-67.

Lynch, A. J., Myers, B. J. E., Chu, C., Eby, L. A., Falke, J. A., Kovach, R. P., Krabbenhoft, T. J., Kwak, T. J., Lyons, J., Paukert, C.P., and Whitney, J.E. 2016. Climate change effects on North American inland fish populations and assemblages. Fisheries. **41**(7):346-361.

Magalhaes, M.F., Beja, P., Schlosser, I.J. and Collares‐Pereira, M.J. 2007. Effects of multi‐year droughts on fish assemblages of seasonally drying Mediterranean streams. Freshwater Biology, **52**(8):1494-1510.

Martin, S.L. and Soranno, P.A. 2006. Lake landscape position: Relationships to hydrological connectivity and landscape features. Limnology & Oceanography, **51**(2), 801-814.

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low‐Choy, S.J., Tyre, A.J. and Possingham, H.P. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters, **8**(11), 1235-1246.

Minns, C.K., and Moore, J.F. 1995 Factors limiting the distributions of Ontario's freshwater fishes: the role of climate and other variables, and the potential impacts of climate change. In: Beamish, R.J. ed., Climate Change and Northern Fish Populations. Canadian Special Publication of Fisheries and Aquatic Sciences 121.

Mittlböck, M. and Waldhör, T. 2000. Adjustments for *R*2-measures for Poisson regression models. Computational Statistics and Data Analysis. **34**(4),461-472.

Muñoz-Mas, R., Gil-Martínez, E., Oliva-Paterna, F.J., Belda, E.J. and Martinez-Capel, F., 2019. Tree-based ensembles unveil the microhabitat suitability for the invasive bleak (*Alburnus alburnus* L.) and pumpkinseed (*Lepomis gibbosus* L.): Introducing XGBoost to eco-informatics. Ecological Informatics, **53**, p.100974.

Myers, B.J.E., Lynch, A.J., Bunnell, D.B., Chu, C., Falke, J.A., Kovach, R.P., Krabbenhoft, T.J., Kwak, T.J., Paukert, C.P. 2017. Global synthesis of the projected and documented effects of climate change on inland fishes. Reviews in Fish Biology and Fisheries, **27**(2):339-361.

Nielsen, D. 2016. Tree boosting with xgboost - why does xgboost win "every" machine learning competition? Master's thesis, Norwegian University of Science and Technology NTNU.

O’Donnell, M.S., and Ignizio, D.A. 2012, Bioclimatic predictors for supporting ecological applications in the conterminous United States: U.S. Geological Survey Data Series 691, 10 pp.

Ouranos. 2015. Vers l’adaptation. Synthèse des connaissances sur les changements climatiques au Québec. Édition 2015. 415 pp, Ouranos Consortium on Regional Climatology and Adaptation to Climate Change, Montréal, Québec.

Pichler, M. and Hartig, F. 2023. Machine learning and deep learning — A review for ecologists. Methods in Ecology and Evolution, **14**(4), 994-1016.

Poesch, M.S., Chavarie, L., Chu, C., Pandit, S.N. and Tonn, W. 2016. Climate change impacts on freshwater fishes: a Canadian perspective. Fisheries, **41**(7):385-391.

Randall, R.G., Minns, C.K. and Kelso, J.R.M. 1995. Fish production in freshwaters: are rivers more productive than lakes?. Canadian Journal of Fisheries and Aquatic Sciences, **52**(3):631-643.

Rayner, T.S., Pusey, B.J. and Pearson, R.G. 2008. Seasonal flooding, instream habitat structure and fish assemblages in the Mulgrave River, north-east Queensland: towards a new conceptual framework for understanding fish-habitat dynamics in small tropical rivers. Marine and Freshwater Research, **59**(2):97-116.

Reid, M.A., Delong, M.D. and Thoms, M.C. 2012. The influence of hydrological connectivity on food web structure in floodplain lakes. River Research and Applications, **28**(7), pp.827-844.

Reist, J.D., Wrona, F.J., Prowse, T.D., Power, M., Dempson, J.B., Beamish, R.J., King, J.R., Carmichael, T.J. and Sawatzky, C.D. 2006. General effects of climate change on Arctic fishes and fish populations. AMBIO: A Journal of the Human Environment, **35**(7):370-380.

Ressources naturelles Canada (2016). CanVec. Ottawa, Canada.

Schindler, D.W. 2001. The cumulative effects of climate warming and other human stresses on Canadian freshwaters in the new millennium. Canadian Journal of Fisheries and Aquatic Sciences, **58**(1), 18-29.

Scott, W.B., and E.J. Crossman. 1973. Freshwater Fishes of Canada. Fisheries Research Board of Cananda, Bulletin **184**:1-966.

Shuter, B.J. and Post, J.R. 1990. Climate, population viability, and the zoogeography of temperate fishes. Transactions of the American Fisheries Society, **119**(2):314-336.

Stoklosa, J., Blakey, R.V. and Hui, F.K., 2022. An overview of modern applications of negative binomial modelling in ecology and biodiversity. Diversity, **14**(5):320.

Sutfin, N.A. and Wohl, E. 2019. Elevational differences in hydrogeomorphic disturbance regime influence sediment residence times within mountain river corridors. Nature Communications, **10**(1):2221.

Suttle, K.B., Power, M.E., Levine, J.M. and McNeely, C. 2004. How fine sediment in riverbeds impairs growth and survival of juvenile salmonids. Ecological Applications, **14**(4), pp.969-974.

Tennant, C.J., Crosby, B.T. and Godsey, S.E. 2015. Elevation‐dependent responses of streamflow to climate warming. Hydrological Processes, **29**(6):991-1001.

Van Zuiden, T.M., Chen, M.M., Stefanoff, S., Lopez, L. and Sharma, S. 2016. Projected impacts of climate change on three freshwater fishes and potential novel competitive interactions. Diversity and Distributions, **22**(5):603-614.

Waldock, C., Stuart‐Smith, R.D., Albouy, C., Cheung, W.W., Edgar, G.J., Mouillot, D., Tjiputra, J. and Pellissier, L. 2022. A quantitative review of abundance‐based species distribution models. Ecography, 2022:e05694.

Warren, D.R., Dunham, J.B. and Hockman-Wert, D. 2014. Geographic variability in elevation and topographic constraints on the distribution of native and nonnative trout in the Great Basin. Transactions of the American Fisheries Society, **143**(1):205-218.

Warren, M., Dunbar, M.J. and Smith, C. 2015. River flow as a determinant of salmonid distribution and abundance: a review. Environmental Biology of Fishes, **98**:1695-1717.

Weigel, D.E. and Sorensen, P.W. 2001. The influence of habitat characteristics on the longitudinal distribution of brook, brown, and rainbow trout in a small Midwestern stream. Journal of Freshwater Ecology, **16**(4), pp.599-613.

Wurtsbaugh, W.A., Heredia, N.A., Laub, B.G., Meredith, C.S., Mohn, H.E., Null, S.E., Pluth, D.A., Roper, B.B., Saunders, W.C., Stevens, D.K. and Walker, R.H. 2015. Approaches for studying fish production: Do river and lake researchers have different perspectives?. Canadian Journal of Fisheries and Aquatic Sciences, 72(1):149-160.

Zhang, C., Chen, Y., Xu, B., Xue, Y. and Ren, Y. 2020. Improving prediction of rare species’ distribution from community data. Scientific Reports, **10**(1):12230.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 1. Thermal traits, habitat preference, and goodness-of-fit statistics of the XGBoost model for the 10 study species, ordered by increasing temperature preference. | | | | | | |
| Species | Latin name | Thermal group a | Thermal preferendum (°C) | Habitat preference b | Pseudo-*R*2 | Cross-validation  pseudo-*R*2 |
| Arctic char | *Salvelinus alpinus* | cold | 9.2 | none | 0.93 | 0.73 |
| lake trout | *Salvelinus namaycush* | cold | 10 | lentic | 0.91 | 0.76 |
| lake whitefish | *Coregonus clupeaformis* | cold | 12.7 | none | 0.88 | 0.66 |
| brook trout | *Salvelinus fontinalis* | cold | 16 | none | 0.87 | 0.69 |
| Atlantic salmon c | *Salmo salar* | cold | 16 | lotic | 0.93 | 0.84 |
| lake sturgeon | *Acipenser fulvescens* | cold/cool | 15-17 | none | 0.93 | 0.86 |
| sauger | *Sander canadensis* | cool | 19.2 | none | 0.98 | 0.93 |
| yellow perch | *Perca flavescens* | cool | 21.4 | lentic | 0.82 | 0.52 |
| walleye | *Sander vitreus* | cool | 22 | lentic | 0.91 | 0.76 |
| northern pike | *Esox lucius* | cool | 22.5 | none | 0.83 | 0.63 |

a From Table 2 in Coker et al. (2001).

b Based on Scott and Crossman (1973).

c Excluding landlocked ("ouananiche") populations.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 2. Predictor variables used in the XGBoost and ZINB models. | | | | | |
| Variable name | Abbreviation | Bioclim code | Mean | 10% quantile | 90% quantile |
| Annual Mean Temperature (°C) | Temp\_Ann\_Mean | BIO1 | 0.63 | -2.46 | 3.91 |
| Mean Diurnal Temperature Range (mean of monthly (max temp - min temp); °C) | Temp\_Diel\_Range | BIO2 | 10.53 | 9.34 | 11.76 |
| Temperature Annual Range (max temperature of warmest month - min temperature of coldest month; °C) | Temp\_Ann\_Range | BIO7 | 45.23 | 41.12 | 47.89 |
| Annual Precipitation (mm) | Precip\_Ann | BIO12 | 943.61 | 687 | 1173 |
| Precipitation Seasonality (coefficient of variation) | Precip\_Season | BIO15 | 30.19 | 16.05 | 46.93 |
| Elevation (m.a.s.l) | Elevation | - | 284.45 | 56 | 495 |
| Habitat (1 = lake; 2 = river) | Habitat | - | 1.48 | 1 | 2 |
| Year (1973 through 2021) | Year | - | 1996.3 | 1976 | 2015 |
| Month (1 through 12) | Month | - | 7.57 | 6 | 9 |
| Gillnet indicator (0 or 1) | Gillnet | - | 0.77 | 0 | 1 |
| Seine indicator (0 or 1) | Seine | - | 0.05 | 0 | 0 |

**FIGURE CAPTIONS**

Figure 1. Sampling sites across the province of Quebec. Colours represent the three different data sources: Hydro-Québec, the former MFFP (Ministère des Forêts, de la Faune et des Parcs du Québec), and the former MELCC (Ministère de l’Environnement et de la Lutte contre les Changements Climatiques du Québec). Map was realized with ArcMap 10.4.1 (Esri) and the 1:5M CanVec database (Ressources naturelles Canada 2016) projected on a Lambert conformal conic projection (datum NAD 1983).

Figure 2. Importance of predictors, as measured by SHAP (SHapley Additive exPlanations) values, for the ten fish species, with the most important predictors at the top of each plot. Predictor name abbreviations as in Table 2. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Climate variables (temperature and precipitation) are indicated in dark blue. Landscape variables are in turquoise: elevation and habitat type (lake or river). Temporal variables associated with sampling are in yellow and gear type is in red.

Figure 3. Importance of predictors across the ten species models. For each species, the importance values of individual predictors (mean of the absolute SHAP values for the predictor; see Materials and Methods) were transformed to ranks to facilitate the comparison of predictors across species. Predictors along the X axis are ordered by decreasing mean value of SHAP rank. Climate variables (temperature and precipitation) are indicated in dark blue. Landscape variables are in turquoise: elevation and habitat type (lake or river). Temporal variables associated with the sampling are in yellow and gear types are in red. Predictor name abbreviations as in Table 2.

Figure 4. Dependence plots for annual mean temperature based on the SHAP values and fitted loess curves. Each SHAP value represents the difference between the Poisson mean predicted for an individual observation and the overall average model prediction. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Each point represents an individual observation (sampling site).

Figure 5. Dependence plots for annual mean precipitation based on the SHAP values and fitted loess curves. Each SHAP value represents the difference between the Poisson mean predicted for an individual observation and the overall average model prediction. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Each point represents an individual observation (sampling site).

Figure 6. Dependence plots for precipitation seasonality based on the SHAP values and fitted loess curves. Each SHAP value represents the difference between the Poisson mean predicted for an individual observation and the overall average model prediction. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Each point represents an individual observation (sampling site).

Figure 7. Dependence plots for elevation based on the SHAP values and fitted loess curves. Each SHAP value represents the difference between the Poisson mean predicted for an individual observation and the overall average model prediction. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Each point represents an individual observation (sampling site).

Figure 8. SHAP boxplots showing the distribution of SHAP values by habitat type (lake or river). The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference.

Figure 9. Metrics of habitat preference for the 10 focal species in our dataset.Top: two preference metrics, the log-ratio of mean abundances in each habitat (red), and the difference in SHAP values of lakes and rivers (black) are shown (see Methods for details). Symbols represent three categories of habitat preference compiled from the literature (lentic: squares; lotic: triangles; none: circles; see Table 1). For both metrics, positive values indicate preference for lakes, negative values preference for rivers; values near zero indicate no habitat preference. The shaded area delimits the SHAP-based values containing all of the species classified as "none" (circles). Bottom: relationship between the two preference metrics and the categories of habitat preference compiled from the literature. Symbols and vertical lines correspond to means ± 1 s.d. Color coding of metrics as in the top plot.

A map of canada with red and blue spots

Description automatically generated

Figure 1. Sampling sites across the province of Quebec. Colours represent the three different data sources: Hydro-Québec, the former MFFP (Ministère des Forêts, de la Faune et des Parcs du Québec), and the former MELCC (Ministère de l’Environnement et de la Lutte contre les Changements Climatiques du Québec).

A group of colorful bars with text

Description automatically generated with medium confidence

Figure 2. Importance of predictors, as measured by SHAP (SHapley Additive exPlanations) values, for the ten fish species, with the most important predictors at the top of each plot. Predictor name abbreviations as in Table 2. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Climate variables (temperature and precipitation) are indicated in dark blue. Landscape variables are in turquoise: elevation and habitat type (lake or river). Temporal variables associated with sampling are in yellow and gear type is in red.

A graph with different colored bars

Description automatically generated

Figure 3. Importance of predictors across the ten species models. For each species, the importance values of individual predictors (mean of the absolute SHAP values for the predictor; see Materials and Methods) were transformed to ranks to facilitate the comparison of predictors across species. Predictors along the X axis are ordered by decreasing mean value of SHAP rank. Climate variables (temperature and precipitation) are indicated in dark blue. Landscape variables are in turquoise: elevation and habitat type (lake or river). Temporal variables associated with the sampling are in yellow and gear types are in red. Predictor name abbreviations as in Table 2.

A graph of different types of fish

Description automatically generated

Figure 4. Dependence plots for annual mean temperature based on the SHAP values and fitted loess curves. Each SHAP value represents the difference between the Poisson mean predicted for an individual observation and the overall average model prediction. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Each point represents an individual observation (sampling site).

A group of blue and red graphs

Description automatically generated

Figure 5. Dependence plots for annual mean precipitation based on the SHAP values and fitted loess curves. Each SHAP value represents the difference between the Poisson mean predicted for an individual observation and the overall average model prediction. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Each point represents an individual observation (sampling site).

A group of blue and red graphs

Description automatically generated

Figure 6. Dependence plots for precipitation seasonality based on the SHAP values and fitted loess curves. Each SHAP value represents the difference between the Poisson mean predicted for an individual observation and the overall average model prediction. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Each point represents an individual observation (sampling site).

A group of graphs showing different types of data

Description automatically generated with medium confidence

Figure 7. Dependence plots for elevation based on the SHAP values and fitted loess curves. Each SHAP value represents the difference between the Poisson mean predicted for an individual observation and the overall average model prediction. The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference. Each point represents an individual observation (sampling site).

​A group of graphs with text

Description automatically generated with medium confidence​

Figure 8. SHAP boxplots showing the distribution of SHAP values by habitat (lake or river). The top row shows responses in the cold-water species (preference <16 °C) and the bottom row shows responses in cold/cool-water (lake sturgeon) and cool-water species (preference ≥ 16 °C), all ordered by increasing temperature preference.

A diagram of different types of fish

Description automatically generated

Figure 9. Metrics of habitat preference for the 10 focal species in our dataset. Top: two preference metrics, the log-ratio of mean catches in each habitat (red), and the difference in SHAP values of lakes and rivers (black) are shown (see Methods for details). Symbols represent three categories of habitat preference compiled from the literature (lentic: squares; lotic: triangles; none: circles; see Table 1). For both metrics, positive values indicate preference for lakes, negative values preference for rivers; values near zero indicate no habitat preference. The shaded area delimits the SHAP-based values containing all the species classified as "none" (circles). Bottom: relationship between the two preference metrics and the categories of habitat preference compiled from the literature. Symbols and vertical lines correspond to means ± 1 s.d. Colour coding of metrics as in the top plot.