Quantifying status and trends from monitoring surveys: application to pygmy whitefish (*Prosopium coulterii*) in Lake Superior

Adam S. van der Lee, Mark R. Vinson, and Marten A. Koops

Abstract: Population assessments of fish species often rely on data from surveys with different objectives, such as measuring biodiversity or community dynamics. These surveys often contain spatial-temporal dependencies that can greatly influence conclusions drawn from analyses. Pygmy whitefish (PWF, *Prosopium coulterii*) populations in Lake Superior were recently assessed as Threatened by the Committee on the Status of Endangered Species in Canada, which motivated a thorough analysis of available data to improve our understanding of its population status. The US Geological Survey conducts annual bottom trawl surveys in Lake Superior that commonly capture PWF. We used these data (1989–2018) to model temporal trends in PWF biomass density and make lake-wide population projections. We used a Bayesian approach, integrated nested Laplace approximation (INLA), and compared the impact of including different random structures on model fit. Inclusion of spatial structure improved model fit and conclusions differed from models omitting random effects. PWF populations have experienced periodic fluctuations in biomass density since 1989, though 2018 may represent the lowest density in the 30-year time series. Lake-wide biomass was estimated to be 71.5 t.

Résumé : Les évaluations de populations d'espèces de poissons reposent souvent sur des données de relevés aux objectifs variés, comme la mesure de la biodiversité ou de la dynamique des communautés. Ces relevés contiennent souvent des dépendances spatiotemporelles qui peuvent grandement influencer les conclusions tirées des analyses. Le Comité sur la situation des espèces en péril au Canada a récemment établi que les populations de ménominis pygmées (*Prosopium coulterii*) dans le lac Supérieur sont menacées, ce qui a motivé une analyse exhaustive des données disponibles dans le but d'améliorer la compréhension du statut des populations de ce poisson. Le U.S. Geological Survey effectue des relevés annuels au chalut de fond dans le lac Supérieur qui capturent généralement des ménominis pygmées. Nous avons utilisé ces données (1989–2018) pour modéliser les tendances dans le temps de la densité de biomasse de ménominis pygmées et établir des projections concernant les populations pour l'ensemble du lac. Nous avons utilisé une approche bayésienne, l'approximation de Laplace emboîtée intégrée (ALEI), et comparé l'incidence de l'inclusion de différentes structures aléatoires sur le calage des modèles. L'inclusion d'une structure spatiale améliore le calage et mène à des conclusions différentes de celles découlant de modèles qui n'intègrent pas d'effets aléatoires. Si la densité de la biomasse de populations de ménominis pygmées a connu des fluctuations périodiques depuis 1989, l'année 2018 pourrait être caractérisée par la plus faible densité de la série chronologique de 30 ans. La biomasse dans l'ensemble du lac est estimée à 71,5 t. [Traduit par la Rédaction]

Introduction

Fish population assessments, whether for the purposes of conservation or fisheries management, must often rely on data from monitoring surveys designed to meet other objectives. For example, monitoring surveys may be designed with a focus on biodiversity or community dynamics, but are also, out of necessity, used to assess single-species population trends and stock assessments. The design of monitoring surveys should reflect the primary objectives and end-use of the data (e.g., Pollock et al. 2002). However, the resource-constrained reality is that fish surveys are often applied to questions and problems other than their original intent. A consequence of this is that the data acquired from monitoring surveys may have both temporal and spatial

dependencies that can complicate the analyses and influence the nature and reliability of inferences.

The US Geological Survey (USGS) conducts an annual bottom trawl survey in Lake Superior to assess the status and trends of the prey fish community (Vinson et al. 2018). Nearshore trawls in United States waters of the lake have been conducted since the early 1960s. In 1989, nearshore locations in Canadian waters were added to the survey. In 2011, an additional offshore survey was begun at deepwater sites (>100 m) in Canada and the United States. While designed as a prey fish community survey, prior analyses have shown that it does provide reliable species level population metrics for some commercially important species, such as lake whitefish (*Coregonus clupeaformis*; Curtis et al. 1993)

Received 9 June 2021. Accepted 5 October 2021.

A.S. van der Lee and M.A. Koops.* Fisheries and Oceans Canada, Great Lakes Laboratory for Fisheries and Aquatic Sciences, 867 Lakeshore Rd., Burlington, ON L7S 1A1, Canada.

M.R. Vinson. US Geological Survey, Great Lakes Science Center, Lake Superior Biological Station, Ashland, WI 54806, USA.

Corresponding author: Adam S. van der Lee (email: adam.vanderlee@dfo-mpo.gc.ca).

*Marten A. Koops served as an Associate Editor at the time of manuscript review and acceptance; peer review and editorial decisions regarding this manuscript were handled by J. Michael Jech.

© 2021 The Crown. This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Fig. 1. Trawl locations and pygmy whitefish (*Prosopium coulterii*) catch values in 2018 from the USGS bottom trawl survey.

and cisco (*Coregonus artedii*; Yule et al. 2006). Potential improvements to the survey have been identified (Yule et al. 2008), but not implemented.

Pygmy whitefish (PWF, *Prosopium coulterii*, maximum size ~150 mm) have a highly discontinuous distribution across North America (COSEWIC 2016). PWF are present in Lake Superior and four additional lakes in northwestern Ontario and at disjunct locations in northwestern United States and western Canada. PWF are frequently caught during the USGS nearshore survey, representing the fifth most abundant species in nearshore trawls. PWF are infrequently caught in the offshore survey (Vinson et al. 2018). PWF populations in Lake Superior were assessed as Threatened by the Committee on the Status of Endangered Wildlife in Canada (COSEWIC) on the basis of an estimated 48% decline in abundance since 2000 in USGS nearshore trawls (COSEWIC 2016). PWF, however, is not a well-studied species, and a detailed analysis of the available catch data has not been conducted. The COSEWIC assessment highlights the need to improve our understanding of PWF in Lake Superior, including an assessment of populations trends and estimating population size.

We used the USGS bottom trawl dataset to estimate the long-term trend in PWF biomass in Lake Superior and to estimate lake-wide population size. Because the same sampling locations are revisited each year, there is a dependency in the data at the site level across years. As well, there is additional within-year spatial dependency as proximate sampling locations are likely to be more similar than locations further apart. The effects of these dependencies can often be accounted for with the inclusion of random effects in statistical models. We implement a hierarchical Bayesian approach, integrated nested Laplace approximation (INLA; Rue et al. 2009), in our analysis. INLA uses deterministic approximations rather than Markov chain Monte Carlo (MCMC) sampling, which provides a method for making fast and accurate Bayesian approximations. In addition, when combined with the stochastic partial differential equation approach (SPDE; Lindgren et al. 2011), INLA can estimate Gaussian Markovian random fields (GMRF) to account for complex covariance structures typical in spatial-temporal data. Use of spatial INLA models is becoming common in marine fisheries analyses (e.g., Cosandey-Godin et al. 2015; Quiroz et al. 2015; Breivik et al. 2017; Rufener et al. 2017; Nikolioudakis et al. 2019; Lezama-Ochoa et al. 2020) but has been applied less frequently in fresh water (e.g., Gutowsky et al. 2019; Enders et al. 2021). We fit three model types with different random structures that either ignored or accounted for the dependencies within the data to assess how well the models fit the data

and to determine how the conclusions derived from the model outputs differed.

Methods

Data

The USGS conducts separate surveys to sample nearshore and offshore locations in Lake Superior (Fig. 1). Nearshore sampling sites were located around the perimeter of the lake with sampling typically occurring in May and June (~3% of trawls took place in April, July, or August) during daylight hours. Trawls were conducted with a 12 m Yankee bottom trawl with either a chain or 15 cm rubber roller foot rope (Vinson et al. 2018). The roller foot rope was used at sites with steeper, rockier bottoms to reduce snagging. A mean of 77 nearshore sites were sampled annually (range: 52–87). Nearshore trawls were conducted across depth contours. Depth was measured at the start and end of each trawl. Start depths ranged from 9 to 68 m (median: 16 m) and end depths ranged from 12 to 144 m (median: 59 m). The median increase in depth within a trawl was 41 m. The mean of start and end depth was used to represent depth in the analyses, with mean depth ranging from 15 to 92 m (median: 39 m).

The offshore survey occurred in July except for 2012 when sampling was conducted in August. A mean of 38 offshore sites were sampled annually (range: 30–53). The variability in the number of samples across years was due to 2011 and 2016 when additional shallower sites (~85 m) were sampled; the number of locations sampled in other years ranged from 30–36 sites. Offshore trawls were conducted along depth contours with sampling depths ranging from 87 to 315 m (median: 181 m).

Trawl collections were counted and weighed by species. Species-specific fish abundance (number·ha⁻¹) and biomass density (kg·ha⁻¹) were estimated by dividing sample counts and weights by the bottom area fished during each trawl.

Statistical analysis

The addition of offshore trawls in 2011 was a significant change to the survey. Therefore, the dataset was split and analysed separately. The first analysis included only nearshore trawls from 1989–2018 with the purpose of identifying long-term trends in occurrence and biomass density of PWF (hereinafter referred to as long-term trend analysis). The second analysis included nearshore and offshore trawls from 2011–2018 with the purpose of estimating lake-wide biomass (hereinafter referred to as lake-wide biomass analysis). The long-term trend dataset consisted of 2314 trawls, of which 949 (41%) caught PWF. The lake-wide

Table 1. Summary of models and model components in the analysis.

Analysis	Model	Fixed effect	Random effects
Long-term trend analysis	GLM	Intercept	Year
	Mixed model	Intercept	Year, site ID
	Spatial model	Intercept	Year, location (correlated)
Lake-wide biomass analysis	GLM	Intercept, depth, depth ² , easting, northing	Year
	Mixed model	Intercept, depth, depth ² , easting, northing	Year, site ID
	Spatial model	Intercept, depth, depth ²	Year, location (correlated)

Note: GLM is generalized linear model.

biomass dataset consisted of 919 trawls, of which 314 (34%) caught PWF.

Due to the high percentage of trawls in which no PWF were caught, we used hurdle models in the analysis. A hurdle model is a two-part model: a Bernoulli process to analyse the presence-absence data and a continuous process (here a gamma model) to analyse the positive catches. Presence-absence, $y_{i,t}^{01}$, at location i and time t was modelled as follows (Zuur and Ieno 2018):

- (1) $y_{i,t}^{01} = \text{Bernoulli}(\pi_{i,t})$
- (2) $E(y_{i,t}^{01}) = \pi_{i,t}$ and $\text{var}(y_{i,t}^{01}) = \pi_{i,t} \times (1 - \pi_{i,t})$
- (3) $\text{logit}(\pi_{i,t}) = Z_{i,t} + u_t + e$

and positive biomass density ($y_{i,t}^{>0}$, kg·ha⁻¹) was modelled as follows (Zuur and Ieno 2018):

- (4) $y_{i,t}^{>0} = \text{Gamma}(\mu_{i,t}, r)$
- (5) $E(y_{i,t}^{>0}) = \mu_{i,t}$ and $\text{var}(y_{i,t}^{>0}) = \mu_{i,t}^2 / r$
- (6) $\log(\mu_{i,t}) = Z_{i,t} + u_t + e$

where π represents mean occurrence in the catch and μ represents the mean biomass density when PWF was present, with r the gamma distribution shape parameter. $Z_{i,t}$ represents all fixed effects (e.g., intercept or intercept and covariates). The trend in time, u_t , was modelled as a random walk function with noise term w_t , such that (Zuur et al. 2017)

- (7) $u_t = u_{t-1} + w_t$

where

$$w_t \sim N(0, \sigma_w^2)$$

Additional random effects incorporated into the models are represented by e . For each analysis, we fit three models that differed in their random structure to account for different levels of dependencies in the data (Table 1). Hyperparameters for the random effects were estimated separately for the presence-absence and positive catch portions of the hurdle models. The simplest model included no additional random structure ($e = 0$) and is referred to as the generalized linear model (GLM). The second model included a random effect for sampling location accounting for the potential dependency at each sampling location across years where $e = \text{site}$ and the random effects are independent and identically distributed with a mean of 0 and variance of

Fig. 2. Temporal trends in pygmy whitefish (*Prosopium coulterii*) catch probability predicted from long-term trend analysis using models with different random structures fit to long-term (1989–2018), nearshore bottom trawl data. [Colour online.]

σ_{site}^2 . This model is referred to as the mixed model. The third model included correlated random effects at sampling locations accounting for potential dependencies at each sampling location across years as well as potential correlations among observations that are closer together in space. This model is referred to as the spatial model. For this model, $e = v_i$, where v_i is a GMRF with mean 0 and covariance matrix Σ . The covariance matrix Σ included variance σ_v^2 and correlations estimated by the Matérn correlation function using SPDE (Lindgren et al. 2011). In INLA, the Matérn correlation function is defined by an estimated range parameter, R , which describes the distance at which the correlation between observations drops to 0.1. The GMRF is estimated over a mesh consisting of non-overlapping triangles constructed with built-in INLA functions (see online Supplement; Fig. S1¹).

The long-term trend analysis consisted of 30 years of data. To allow for possible shifts in spatial distribution over time, multiple spatial fields were estimated. Rather than estimate a spatial field for each year of data, which is computationally intensive, four spatial fields were estimated at evenly distributed knots, k , in the time series (1992, 2000, 2008, and 2016). At each knot a spatial field was estimated as the weighted average of data from the surrounding years. The knot years were selected such that each spatial field had an approximately equal weight (total number of years in time series influenced by each spatial field). Correlation between spatial fields was assumed following an AR1 process with correlation ϕ ; therefore, the spatial field was a function of the previous field and the new spatial effect, $s_{i,k}$, such that the random spatial effect becomes (Krainski et al. 2019)

¹Supplementary data are available with the article at <https://doi.org/10.1139/cjfas-2021-0155>.

Fig. 3. Temporal trends in pygmy whitefish (*Prosopium coulterii*) mean biomass density ($\text{kg}\cdot\text{ha}^{-1}$) predicted from long-term trend analysis using models with different random structures fit to long-term (1989–2018), nearshore bottom trawl data. [Colour online.]

$$(8) \quad v_{i,k} = \Phi v_{i,k-1} + s_{i,k}$$

The lake-wide biomass analysis represents only eight years of data and therefore a single spatial field was estimated for the model assuming that PWF distribution did not change over that time.

In the long-term trend analysis, we were only interested in the temporal trend in occurrence and biomass density, so additional covariates were not incorporated into the models. In the lake-wide biomass analysis, additional covariates were necessary to make lake-wide projections (Table 1). We included depth as a second-degree polynomial in all models. Preliminary analysis revealed a nonlinear effect of depth on occurrence and biomass density with the second-degree polynomial providing a superior fit to higher degree polynomials. As well, positional data (as eastings and northings (km)) were included as continuous linear variables in the GLM and mixed model. Positional data were not included as fixed effects in the spatial model as this information is represented in the spatial field. Vague priors were used for all model components.

The lake-wide biomass analysis models were used to make spatial projections of PWF biomass for 2018 based on the probability of occurrence and biomass density predictions. The projections were made on a 1 km^2 grid across the lake. The projections were summed to give estimates of lake-wide population size. Occupied habitat and preferred habitat were estimated as the amount of habitat with projected biomass density $>0 \text{ kg}\cdot\text{ha}^{-1}$ and the amount of habitat with $>50\%$ predicted occurrence, respectively. To incorporate random effects from the mixed model the values were interpolated using inverse distance weighting and the package *intamap* (Pebesma et al. 2011).

Model comparison

The goodness-of-fit of the models with different random structures were compared across both analyses using multiple methods. Overall model fit was measured with WAIC (widely applicable information criterion, Watanabe 2010). We also assessed prediction accuracy for both the presence-absence and positive biomass density aspects of the hurdle models. The effectiveness of the presence-absence model was tested by examining the in-sample prediction accuracy comparing observed to fitted values, where predicted catch probability >0.5 was taken as a prediction of presence. The accuracy of the positive biomass density portion of the model was assessed by correlating the in-sample fitted values with the observed values and calculating the root mean squared estimation error (RMSEE) calculated as the square root of the squared residuals divided by number of observations:

$$(9) \quad \text{RMSEE} = \sqrt{\frac{1}{n} \sum (y - \hat{y})^2}$$

Results

The long-term trend analysis found no evidence of a significant trend in PWF catch probability through time regardless of model fit (Fig. 2), and the standard deviation associated with the year effect was low (<0.1) for each model. The estimate of mean occurrence in the catch across years was highest with the GLM (0.41) and similar for the mixed and spatial models (0.32 and 0.28, respectively). Changes in PWF biomass density through time were more significant. Error associated with the year trend was higher for the GLM model (0.28) than for the mixed (0.20) or spatial model (0.22), and there were greater differences in estimated trends among the models (Fig. 3). The GLM model indicated strong intradecadal variation, but a continuous decline in biomass density since the early 1990s. The trends from the mixed and spatial model, however, were similar to each other and were characterized as intradecadal fluctuations in biomass through the study period. Although both models indicated a decline in biomass density in recent years (since 2013), the spatial model estimated mean biomass density in 2018 to be at the lowest point in the 30-year time series ($<0.02 \text{ kg}\cdot\text{ha}^{-1}$). The trend from the GLM differed because it is biased towards large catch events, which are rare (for example catches $>3 \text{ kg}\cdot\text{ha}^{-1}$ occurred in 17 trawls across 6 sampling locations). The frequency of large catch events has decreased in recent years (the most recent sample $>3 \text{ kg}\cdot\text{ha}^{-1}$ was in 2014), suggesting a potential decline in biomass among high density subpopulations; however, this is somewhat confounded with changes in sampling effort (2 of the 6 high-density sites have not been sampled since 2003 and 2013). The random effects structures incorporated into the mixed and spatial models reduces the bias from these sites and gives a better representation of the average trend across the lake.

The random effects parameter estimates appear important. The standard deviation associated with sampling location in the mixed model was 0.13 for the presence-absence portion of the model and 0.43 for the positive biomass portion of the model. The standard deviation for the spatial fields were 3.56 and 1.73 for the presence-absence and biomass models. The range estimates were 117 and 60 km indicating at least a 10% correlation between sampling sites up to 117 km apart for presence-absence and 60 km for biomass density. The spatial fields (see Supplement; Figs. S2–S5¹) fit at each knot were highly correlated with Φ estimates of 0.97 and 0.86 for the presence-absences and biomass density components, indicating that PWF spatial occurrence was

Table 2. Goodness-of-fit statistics comparing models with different random structures for the long-term trend and lake-wide biomass analyses.

Analysis	Model	WAIC	Occurrence accuracy	Biomass correlation	RMSEE
Long-term trend analysis	GLM	1089.5	0.59	0.15	0.91
	Mixed model	-1557.3	0.84	0.61	0.73
	Spatial model	-1815.0	0.85	0.71	0.66
Lake-wide biomass analysis	GLM	-163.6	0.77	0.39	0.40
	Mixed model	-502.7	0.87	0.77	0.28
	Spatial model	-531.1	0.86	0.75	0.29

Note: GLM is generalized linear model, WAIC is widely applicable information criterion, and RMSEE is root mean squared estimation error.

Table 3. Hurdle model parameters estimates (median, lower credible interval (LCI), and upper credible interval (UCI)) from the lake-wide biomass analysis for models with different random structures fit to nearshore and offshore bottom trawl data for years 2011–2018.

	Presence-absence			Biomass		
	Median	LCI	UCI	Median	LCI	UCI
GLM						
Fixed effects						
Intercept	1.82	1.36	2.30	-1.05	-1.53	-1.64
Depth	0.30	-0.14	0.71	-0.24	-0.71	0.27
Depth ²	-5.44	-6.50	-4.45	-5.42	-6.47	-4.31
Easting	0.21	0.03	0.39	0.22	0.00	0.45
Northings	0.32	0.15	0.49	0.69	0.46	0.92
Hyperparameters						
σ_e	NA	NA	NA	1.42	1.33	1.52
σ_w	0.15	0.06	0.37	0.15	0.04	0.42
Mixed model						
Fixed effects						
Intercept	1.16	0.14	2.25	-2.28	-2.91	-1.66
Depth	-0.16	-1.22	0.88	0.07	-0.70	0.84
Depth ²	-5.71	-7.90	-3.85	-3.99	-5.30	-2.60
Easting	0.77	0.22	1.33	0.07	-0.29	0.44
Northings	0.33	-0.21	0.44	0.58	0.19	0.96
Hyperparameters						
σ_e	NA	NA	NA	1.08	1.03	1.12
σ_w	0.07	0.06	0.07	0.19	0.15	0.21
σ_{site}	2.76	2.63	2.89	1.54	1.46	1.63
Spatial model						
Fixed effects						
Intercept	1.09	0.12	1.98	-2.65	-3.37	-1.98
Depth	-0.11	-0.86	0.60	0.29	-0.40	1.00
Depth ²	-4.61	-6.00	-3.39	-3.34	-4.52	-2.08
Hyperparameters						
σ_e	NA	NA	NA	1.03	0.95	1.11
σ_w	0.23	0.11	0.48	0.25	0.12	0.51
σ_v	1.81	1.31	2.51	1.53	1.20	1.96
Range	75.75	47.38	127.9	47.99	30.92	76.36

Note: σ_e represents the error associated with the gamma model output observations, σ_w represents the error associated with the year effect, σ_{site} represents the error associated with sampling location, σ_v represents the error associated with the spatial field, and range represents the distance between observations where the correlation is $\geq 10\%$. GLM is generalized linear model.

largely consistent over the study period with only slight changes in the distribution of biomass density.

The goodness-of-fit statistics (Table 2) indicate that the spatial model best fit the data. Across all measures, the spatial model performed the best. The mixed model, however, performed almost

Fig. 4. Trends in pygmy whitefish (*Prosopium coulterii*) catch probability with depth from lake-wide biomass analysis using model with different random structures fit to nearshore and offshore bottom trawl data for years 2011–2018. [Colour online.]

as well in predicting presence-absence as the spatial model. In addition, we examined model residuals to assess violations of model assumptions. The GLM produced large positive residuals, significant autocorrelation based on estimation of the autocorrelation factor, and significant clustering of positive and negative residuals in space. Inclusion of additional random effects at the site and location level reduced this with no apparent violations of model assumptions for the mixed or spatial models.

In the lake-wide biomass analysis, covariates were incorporated into the models (Table 3). Depth as a second-degree polynomial was important across all models. For the GLM, positional data were important for both the presence-absence and biomass density model components with occurrence and biomass density increasing towards the northeastern (Canadian) area of the lake. For the mixed model, only eastings were important in predicting occurrence and only northings were important in predicting biomass density. The model predicts that occurrence increases towards the east side of the lake and biomass density increases towards the Canadian side of the lake. The range value estimates for the spatial model were smaller than in the long-term trend analysis, 76 km for presence-absence and 48 km for biomass density.

Depth was important in predicting PWF catch probability and biomass density. The shape of the trend in catch probability were similar across model fits (Fig. 4). The GLM predicted occurrence to be $>50\%$ between 42 and 126 m depth, while both the mixed

Fig. 5. Trends in mean pygmy whitefish (*Prosopium coulterii*) biomass density ($\text{kg}\cdot\text{ha}^{-1}$) with depth from lake-wide biomass analysis using model with different random structures fit to nearshore and offshore bottom trawl data for years 2011–2018. [Colour online.]

and spatial models predicted occurrence to be $>50\%$ between 46 and 115 m. The GLM estimated a greater maximum occurrence of $\sim 85\%$ between 78 and 90 m depth, while the mixed and spatial models estimated maximum occurrence probabilities of $\sim 74\%$ and $\sim 73\%$ between 75 and 87 m, respectively. Occurrence was generally low ($\sim 5\%$) at depths <15 and >120 m. There were greater differences in the trends of biomass density with depth between the model fits (Fig. 5). The GLM predicted greater mean biomass density at depth than the two other models and the mixed model predicted slightly greater biomass density at depth than the spatial model. The GLM predicted maximum biomass density of $\sim 0.26 \text{ kg}\cdot\text{ha}^{-1}$ at ~ 82.5 m, the mixed model predicted maximum biomass density of $\sim 0.1 \text{ kg}\cdot\text{ha}^{-1}$ at ~ 82.5 m, and the spatial model predicted maximum biomass density of $\sim 0.07 \text{ kg}\cdot\text{ha}^{-1}$ at ~ 85 m.

The GLM fit again performed the worst across all goodness-of-fit measures (Table 2). The spatial model produced the best fit according to WAIC and the mixed model slightly outperformed the spatial model in occurrence accuracy, biomass correlation, and RMSEE. The GLM did provide relatively accurate estimates of occurrence (77%) demonstrating that depth is a strong predictor of PWF presence-absence in Lake Superior. Although the GLM gave poor predictions of biomass density relative to the other models. Model diagnostics were similar to the long-term model. The GLM had large positive residuals and clustering in space, while there were no concerns for the mixed or spatial models.

We used the hurdle models to make spatial prediction of PWF population size based on the likelihood of occurrence and expected biomass density across a 1 km^2 grid of Lake Superior (Fig. 6; Table 4). The GLM predicted the greatest total biomass, mean biomass density, occupied habitat, and preferred habitat. The mixed model predicted the smallest population size while the spatial model predicted the smallest spatial distribution of PWF. The spatial model, however, produce broad credible intervals. The location of projected PWF aggregations differed among the models (Fig. 6). The GLM predicted large aggregations ($>3 \text{ kg}\cdot\text{ha}^{-1}$) along the north shore of the lake. The largest aggregations predicted from the spatial model were in the northeast region of the lake, around Michipicoten Island, with a maximum density of $0.65 \text{ kg}\cdot\text{ha}^{-1}$. The mixed model predicted a lower maximum density ($0.40 \text{ kg}\cdot\text{ha}^{-1}$) with a spatial distribution more similar to the spatial model.

Fig. 6. Spatially explicit predictions of pygmy whitefish (*Prosopium coulterii*) population size (biomass density, $\text{kg}\cdot\text{ha}^{-1}$) for 2018 from lake-wide biomass analysis using models with different random structures.

Discussion

With the COSEWIC (2016) assessment of PWF in Lake Superior as Threatened, a detailed analysis of population trends and abundance is important. We analyzed the available data, a long-term USGS bottom trawl survey, by comparing and contrasting various model fits differing in random structure. The addition of random structure to the model always improved fit and prediction accuracy with the GLM performing the worst across all goodness-of-fit metrics. In analyses of spatial-temporal data, omitting the spatial field has been consistently shown to reduce model performance (Quiroz et al. 2015; Nikolioudakis et al. 2019). The spatial model performed the best in the long-term trend analysis where additional covariates outside of the temporal trend were not considered. The spatial model in the long-term trend analysis had the benefit of allowing the spatial field to change at four knots,

Table 4. Population estimates for pygmy whitefish (*Prosopium coulterii*) in Lake Superior in 2018 based on median (credible interval) predictions from lake-wide biomass hurdle models with different random structures.

	GLM	Mixed model	Spatial model
Total biomass (kg)	279 156 (146 120–543 069)	56 303 (19 012–159 805)	71 500 (8 560–714 519)
Mean biomass density (kg·ha ⁻¹)	0.088 (0.052–0.156)	0.024 (0.016–0.051)	0.041 (0.021–0.189)
Occupied habitat (km ²)	31 543 (28 358–34 796)	22 696 (11 774–31 020)	17 626 (4 142–37 726)
Preferred habitat (km ²)	13 232 (8 066–17 123)	10 360 (2 684–19 805)	9 162 (4 142–25 170)

Note: Total biomass represents lake-wide population size. Mean biomass density is the average biomass per hectare where pygmy whitefish occur. In occupied habitat, biomass > 0 kg·ha⁻¹. In preferred habitat, occurrence probability > 0.50. GLM is generalized linear model.

whereas the random effects of the mixed model were held constant. In the lake-wide biomass analysis, however, when additional positional and depth covariates were included, the spatial model produced the best fit to the data based on WAIC, but there was little difference between the spatial and mixed models in prediction accuracy. The lack of difference between the mixed and spatial model in prediction accuracy was likely due to large distances between most sampling locations. Across years, only 11.4% of trawl locations were within the 76 km range estimate for the presence-absence portion of the hurdle model, and only 5.4% of trawls were within the 48 km range estimate from the biomass density portion of the hurdle model. Therefore, there was only a small degree of correlation among observations not accounted for in the mixed model.

There were marked differences in the fitted trend through time in relative biomass density (kg·ha⁻¹) between the GLM and mixed or spatial model fits. The GLM showed an almost continuous decline in density through time similar to the conclusions of the COSEWIC assessment (COSEWIC 2016). Alternatively, the mixed and spatial models showed no evidence of a consistent decline but rather periodic cycling in biomass density. As the mixed and spatial models provided a much better fit to the data than the GLM, the trends estimated from these models likely provide a better depiction of the lake-wide trends in PWF density. The results of the GLM are biased towards the sites with the greatest catches, and although they likely are not representative of a lake-wide trend, they demonstrate that there has been a decline in the frequency of large catches at these sites. This may be an indication of a noteworthy decline in what were the strongest PWF subpopulations. The cause of a decrease of large catches in the annual survey is unclear, but it is possibly the result of small shifts in sampling effort. The trend in biomass density from the spatial model shows a decline since 2013 and estimates current densities to be at the lowest levels in the 30-year time series. With this result and the decline in large catches, it will be important to continue monitoring Lake Superior PWF populations into the future.

We provide lake-wide estimates of populations based on the three model types. We have assumed perfect catchability with the bottom trawl, and therefore our population estimates likely underestimate actual population size. Our analysis indicates that the spatial model is the best approach for estimating the spatial distribution of PWF across the lake and total population size. The GLM is biased towards areas of largest catch and overestimates biomass in other locations. It is difficult to project the mixed model across the lake, as there is no method of incorporating the random site effects in space outside of sampling locations. We used an interpolation method, inverse distance weighting; however, it is unlikely this adequately accounted for the underlying spatial distribution of PWF. The spatial field produced from the spatial model and the Matérn correlation function allows for the random effects to be easily projected across the lake and inclusion of all elements of the model in the population projections. The spatial projection INLA produces has multiple applications in fisheries science; for example, they have been used to identify

locations where bycatch risk for Greenland shark (*Somniosus microcephalus*) is elevated (Cosandey-Godin et al. 2015), and areas where the spinytail devil ray (*Mobular mobular*) is expected to occur (Lezama-Ochoa et al. 2020). We recommend use of the spatial model when estimating population size. We project the current population size to be ~71.5 t although with broad credible intervals (CI: 8560–714 519 t), due to the highly right skewed posterior distributions of predictions.

Depth was an important covariate in predicting PWF occurrence and biomass density, both as a second-degree polynomial. Based on the data used in the spatial model, where collection depth was defined as the mean of the trawl start and end depths, PWF were likely (>50%) to occur at depths between 46 and 115 m and were most abundant between ~80 and 95 m. PWF has previously been reported to occur between ~15 and 95 m (Eschmeyer and Bailey 1955; Dryer 1966; Selegby and Hoff 1996). Greatest abundance has been reported at depths slightly shallower than our model results, 55 to 62 m (Eschmeyer and Bailey 1955), 55 to 72 m (Dryer 1966), and 60 m (Yule et al. 2008). These results, however, do not represent whole-lake sampling and were based on abundance rather than biomass.

The relationship of density with depth may differ between biomass and abundance due to potential shifts in depth preferences with body size or age. Younger PWF potentially occupy shallower depths (Eschmeyer and Bailey 1955; Gorman et al. 2012). Eschmeyer and Bailey (1955) and Yule et al. (2008) noted an increase in mean length of PWF with depth. As well, all PWF collected between ~18 and 26 m were young-of-the-year (YOY; Eschmeyer and Bailey 1955; Stewart et al. 2016). The proportion of YOY decreased to 7% beyond 73 m and 0% at 80 m depth (Eschmeyer and Bailey 1955). Our analysis did not incorporate body size or age. Our use of biomass density as the response variable, however, would reduce the influence of young, small fish on the fitted trend and reflect the depth preference of adult PWF in Lake Superior.

Our results were based on cross-contour sampling where the exact depth of capture was unknown. The median increase in depth within a trawl was 41.1 m. Our use of the mean of start and end depths to represent sampling depth in the models adds error to our estimates of depth preferences, potentially broadening the range of preferred depth, leading to overestimates of habitat and population size. Our use of mean depth, rather than taking the start or end depth, allows the error to occur at both tails of the depth distribution preventing a consistent directional bias. A limitation of cross-contour trawling in the nearshore survey is that it does not allow for determining exact depth of capture, which may be important for more accurately documenting species depth occurrences or species distribution shifts due to changing environmental factors, such as climate change (e.g., Dulvy et al. 2008). In the 1970s when the nearshore survey was designed, describing species depth distributions was not the priority. At that time, the focus was on describing the spatial distribution of species' abundance and biomass across the lake for use in stock assessments, and cross-contour trawling was the most efficient way of collecting the most species, with different depth preferences, in a single trawl tow. Conducting multiple on-contour

trawls at specific depths would have reduced the number of sites that were sampled around the lake due the increase in the number of trawls made at each location.

In conclusion, the temporal and spatial dependencies present in the monitoring survey data were informative to the conclusions drawn based on the model applied to the analysis. The addition of random effects provided better fit and greater prediction accuracy to our models of PWF temporal trends and lake-wide biomass. The simple mixed effects model provided trends and a level of accuracy similar to the more complex spatial model as sampling locations were well-spaced relative to estimates of correlation range. The methods associated with the spatial model provided a simpler and more accurate way to make spatial projections and ultimately was the preferred model. We recommend employing these methods when assessing temporal trends of other species captured with bottom trawl surveys in the Great Lakes. The outputs of the spatial model provide an increased understanding of the status of PWF in Lake Superior.

Competing interests

The authors declare there are no competing interests.

Funding

The authors declare no specific funding for this work.

Data availability

USGS Lake Superior bottom trawl research survey data are available online: <https://www.sciencebase.gov/catalog/item/57e185c8e4b0908250033a54>.

References

- Breivik, O.N., Storvik, G., and Nedreaas, K. 2017. Latent Gaussian models to predict historical bycatch in commercial fishery. *Fish. Res.* **185**: 62–72. doi:10.1016/j.fishres.2016.09.033.
- Cosandey-Godin, A., Krainski, E.T., Worm, B., and Flemming, J.M. 2015. Applying Bayesian spatiotemporal models to fisheries bycatch in the Canadian Arctic. *Can. J. Fish. Aquat. Sci.* **72**(2): 186–197. doi:10.1139/cjfas-2014-0159.
- COSEWIC. 2016. COSEWIC assessment and status report on the Pygmy Whitefish *Prosopium coulterii*, Southwestern Yukon Beringian populations, Yukon River populations, Pacific populations, Western Arctic populations, Great Lakes – Upper St. Lawrence populations, Water Committee on the Status of Endangered Wildlife in Canada, Ottawa, Ont.
- Curtis, G.L., Bronte, C.R., and Selgeby, J.H. 1993. Forecasting contributions of lake whitefish year-classes to a Lake Superior commercial fishery from estimates of yearling abundance. *N. Am. J. Fish. Manage.* **13**(2): 349–352. doi:10.1577/1548-8675(1993)013<0349:FCOLWY>2.3.CO;2.
- Dryer, W.R. 1966. Bathymetric distribution of fish in the Apostle Islands Region, Lake Superior. *Trans. Am. Fish. Soc.* **95**(3): 248–259. doi:10.1577/1548-8659(1966)95[248:BDOFTI]2.0.CO;2.
- Dulvy, N.K., Rogers, S.I., Jennings, S., Stelzenmuller, V., Dye, S.R., and Skjoldal, H.R. 2008. Climate change and deepening of the North Sea fish assemblage: a biotic indicator of warming seas. *J. Appl. Ecol.* **45**(4): 1029–1039. doi:10.1111/j.1365-2664.2008.01488.x.
- Enders, E.C., Charles, C., van der Lee, A.S., and Lumb, C.E. 2021. Temporal variations in the pelagic fish community of Lake Winnipeg from 2002 to 2019. *J. Great Lakes Res.* **47**(3): 626–634. doi:10.1016/j.jglr.2021.01.004.
- Eschmeyer, P.H., and Bailey, R.M. 1955. The Pygmy Whitefish, *Coregonus coulteri*, in Lake Superior. *Trans. Am. Fish. Soc.* **84**: 161–199. doi:10.1577/1548-8659(1954)84[161:TPWCCF]2.0.CO;2.
- Gorman, O.T., Yule, D.L., and Stockwell, J.D. 2012. Habitat use by fishes of Lake Superior. I. Diel patterns of habitat use in nearshore and offshore waters of the Apostle Islands region. *Aquat. Ecosyst. Health Manage.* **15**(3): 333–354. doi:10.1080/14634988.2012.715972.
- Gutowsky, L.F.G., Giacomini, H.C., de Kerckhove, D.T., Mackereth, R., McCormick, D., and Chu, C. 2019. Quantifying multiple pressure interactions affecting populations of a recreationally and commercially important freshwater fish. *Global Change Biol.* **25**(3): 1049–1062. doi:10.1111/gcb.14556. PMID: 30580472.
- Krainski, E.T., Gomez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., et al. 2019. Advanced spatial model with stochastic partial differential equations using R and INLA. Chapman & Hall/CRC Press, Boca Raton, Fla.
- Lezama-Ochoa, N., Pennino, M.G., Hall, M.A., Lopez, J., and Murua, H. 2020. Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of the Spinetail Devil Ray (*Mobular mobular*). *Sci. Rep.* **10**(1): 18822. doi:10.1038/s41598-020-73879-3. PMID: 33139744.
- Lindgren, F., Rue, H., and Lindström, J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**(4): 423–498. doi:10.1111/j.1467-9868.2011.00777.x.
- Nikolioudakis, N., Skaug, H.J., Olafsdottir, A.H., Jansen, T., Jacobsen, J.A., and Enberg, K. 2019. Drivers of the summer-distribution of Northeast Atlantic mackerel (*Scomber scombrus*) in the Nordic Seas from 2011 to 2017: A Bayesian hierarchical modelling approach. *ICES J. Mar. Sci.* **76**(2): 530–548. doi:10.1093/icesjms/fsy085.
- Pebesma, E., Cornford, D., Dubois, G., Heuvelink, G.B.M., Hristopulos, D., Pilz, J., et al. 2011. INTAMAP: The design and implementation of an interoperable automated interpolation web service. *Comput. Geosci.* **37**(3): 343–352. doi:10.1016/j.cageo.2010.03.019.
- Pollock, K.H., Nichols, J.D., Simons, T.R., Farnsworth, G.L., Bailey, L.L., and Sauer, J.R. 2002. Large scale wildlife monitoring studies: Statistical methods for design and analysis. *Environmetrics*, **13**(2): 105–119. doi:10.1002/env.514.
- Quiroz, Z.C., Prates, M.O., and Rue, H. 2015. A Bayesian approach to estimate the biomass of anchovies off the coast of Perú. *Biom.* **71**(1): 208–217. doi:10.1111/biom.12227.
- Rue, H., Martino, S., and Chopin, N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71**(2): 319–392. doi:10.1111/j.1467-9868.2008.00700.x.
- Rufener, M.C., Kinas, P.G., Nóbrega, M.F., and Lins Oliveira, J.E. 2017. Bayesian spatial predictive models for data-poor fisheries. *Ecol. Modell.* **348**: 125–134. doi:10.1016/j.ecolmodel.2017.01.022.
- Selgeby, J.H., and Hoff, M.H. 1996. Seasonal bathymetry distributions of 16 fishes in Lake Superior, 1959–75. National Biological Service Biological Science Report 7.
- Stewart, T.R., Ogle, D.H., Gorman, O.T., and Vinson, M.R. 2016. Age, growth, and size of Lake Superior Pygmy Whitefish (*Prosopium coulterii*). *Am. Midl. Nat.* **175**(1): 24–36. doi:10.1674/amid-175-01-24-36.1.
- Vinson, M.R., Evrard, L.M., Gorman, O.T., and Yule, D.L. 2018. Status and trends in the Lake Superior fish community, 2017. In *Compiled reports to the Great Lakes Fishery Commission of the annual bottom trawl and acoustics surveys, 2017*. Great Lakes Fishery Commission. Ann Arbor, Michigan. pp. 1–11.
- Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**: 3571–3594.
- Yule, D.L., Stockwell, J.D., Cholwek, G.A., Evrard, L.M., Schram, S., Seider, M., and Symbal, M. 2006. Evaluation of methods to estimate lake herring spawner abundance in Lake Superior. *Trans. Am. Fish. Soc.* **135**(3): 680–694. doi:10.1577/T05-203.1.
- Yule, D.L., Adams, J.V., Stockwell, J.D., and Gorman, O.T. 2008. Factors affecting bottom trawl catches: implications for monitoring the fishes of Lake Superior. *North Am. J. Fish. Manage.* **28**(1): 109–122. doi:10.1577/M06-289.1.
- Zuur, A.F., and Ieno, E.N. 2018. Beginner's guide to spatial, temporal, and spatial-temporal ecological data analysis with R-INLA. Volume II: GAM and zero-inflated models. Highland Statistics Ltd., Newburgh, UK.
- Zuur, A.F., Ieno, E.N., and Saveliev, A.A. 2017. Beginner's guide to spatial, temporal, and spatial-temporal ecological data analysis with R-INLA. Volume I: Using GLM and GLMM. Highland Statistics Ltd., Newburgh, UK.