Influence of environmental variables on the occurrence of brook trout (Salvelinus fontinalis) within streams of the Lake Simcoe watershed
Influence of environmental variables on the occurrence of brook trout (*Salvelinus fontinalis*) within streams of the Lake Simcoe watershed

Michael W. Thorn¹, Cindy Chu², Nicholas E. Jones²

¹Western University, Department of Biology ²Aquatic Research and Monitoring Section
Some of the information in this document may not be compatible with assistive technologies. If you need any of the information in an alternate format, please contact info.mnrfscience@ontario.ca

Cover photo: Ganaraska River brook trout
Credit: Michael Leung

Cite this report as:
Summary

Brook trout (*Salvelinus fontinalis*) are experiencing population declines throughout their native range due to a variety of human disturbances. The Lake Simcoe watershed is an area that has been impacted by increasing human development and efforts are now being made to maintain and restore the existing brook trout populations. However, there is lack of information regarding the current distribution of brook trout in the Lake Simcoe watershed and this information is needed to direct conservation and management efforts.

Using electrofishing and landscape data, we developed multiple models to predict the distribution of brook trout in the Lake Simcoe watershed and to determine what environmental variables influence their occurrence. The predictions from the distribution models were also combined into an ensemble prediction. Past brook trout distribution was then hindcasted by artificially maximizing the amount of forest in the watershed and reducing human development to zero. The performance of the distribution models was assessed using 5-fold cross-validation and a variety of evaluation metrics including the sensitivity, specificity, true skill statistic, and area under the receiver operating characteristic curve.

All the distribution models performed well and provided accurate predictions of the current and known distribution of brook trout in the Lake Simcoe watershed. The occurrence of brook trout was positively related to overburden thickness, base-flow index, channel slope, stream power index, elevation, and treed area, whereas brook trout were negatively related to the maximum monthly air temperature. Brook trout were restricted mostly to headwater regions. However, there were some streams where the majority of the catchment was suitable for brook trout, including Bluff’s Creek, Kidd’s Creek, and Lovers Creek. The hindcasted distribution was 47% larger than the current distribution and revealed that brook trout have retreated into the headwater regions of the watershed. The results of this study can be used to inform brook trout conservation and restoration efforts in the Lake Simcoe watershed. We also propose several recommendations to improve the monitoring and management of brook trout in the watershed:

- Conserve cold headwater areas where brook trout are present and restore marginal habitat (reforestation and shade riparian) to increase available brook trout habitat and improve connectivity.
- More long-term monitoring sites are required to better assess changes in brook trout population dynamics. Monitoring sites should be located along a gradient of brook trout habitat suitability because this will help in the early identification of distribution changes.
- Sampling sites should be selected using a systematic sampling design, such as stratified random sampling (e.g., aquatic classification).
- Improve electrofishing records (information management) to allow for the estimation of brook trout abundance at all sites by recording accurate shocker seconds and consistently collecting information on the stream area sampled.
- The sampling methodology used to sample each site should be recorded to ensure all data is collected using a standard approach and to indicate when a site was sampled using an alternative methodology.
- Record the life-stage of the fish captured when taking individual and bulk measurements. This will allow the data to be easily separated by life-stage.
- Sample streams during the winter season to provide information on the winter distribution of brook trout in the system.
- All dams and barriers need to be mapped in the watershed, and the degree of connectivity each barrier provides assessed.
- The distribution models could be coupled with groundwater discharge and stream temperature models to better predict areas of occupation.
- The predictions of the distribution models need to be field validated to properly assess their effectiveness.
Résumé

Influence des variables environnementales sur la présence de l’ombre de fontaine (Salvelinus fontinalis) dans les cours d’eau du bassin hydrologique du lac Simcoe

L’ombre de fontaine (Salvelinus fontinalis) connaît un déclin de ses populations dans toute son aire de répartition indigène en raison de diverses perturbations humaines. Le bassin hydrologique du lac Simcoe est une région touchée par l’augmentation du développement humain, et des efforts sont en cours pour maintenir et rétablir les populations existantes d’ombre de fontaine. On manque néanmoins d’information sur la répartition actuelle de l’ombre de fontaine dans le bassin hydrologique du lac Simcoe, alors que cette information est nécessaire pour orienter les efforts de conservation et de gestion.

En nous appuyant sur des données relatives aux paysages et l’électropêche, nous avons mis au point plusieurs modèles pour prédire la répartition de l’ombre de fontaine dans le bassin hydrologique du lac Simcoe et définir les variables environnementales influant sur sa présence. Les prédictions issues des modèles de répartition ont également été combinées dans une prévision d’ensemble. La répartition passée de l’ombre de fontaine a ensuite été simulée en maximisant artificiellement la surface de forêt dans le bassin hydrologique et en réduisant à zéro le développement humain. La performance des modèles de répartition a été évaluée par une validation croisée à cinq volets et par diverses mesures d’évaluation (sensibilité, spécificité, statistiques sur la qualité vraie, zone sous la courbe d’efficacité du récepteur, etc.).

Tous les modèles de répartition ont bien fonctionné et ont fourni des prédictions exactes de la répartition actuelle et connue de l’ombre de fontaine dans le bassin hydrologique du lac Simcoe. La présence de l’ombre de fontaine était en corrélation positive avec l’épaisseur des morts-terrains, l’indice de débit de base, la pente du lit, l’indice de puissance du ruisseau, l’élévation et l’existence d’une zone arborée, mais était en corrélation négative avec la température mensuelle maximale de l’air. L’ombre de fontaine se limitait principalement aux cours supérieurs. Toutefois, dans certains cours d’eau (comme les ruisseaux Bluff’s, Kidd’s et Lovers), la majeure partie du bassin versant convenait à l’ombre de fontaine. La répartition simulée était supérieure de 47 % à la répartition actuelle et révélait que l’ombre de fontaine s’était retiré dans les cours supérieurs du bassin hydrologique. Les résultats de cette étude peuvent servir à orienter les efforts de conservation et de rétablissement de l’ombre de fontaine dans le bassin hydrologique du lac Simcoe. Nous proposons également plusieurs recommandations afin d’améliorer la surveillance et la gestion de cette espèce dans le bassin hydrologique :

- Protéger les eaux froides d’amont où est présente l’ombre de fontaine et rétablir l’habitat marginal (re constitution et plantation de zones riveraines ombragées) afin d’élargir l’habitat convenant à l’ombre de fontaine et de renforcer la connectivité.
- Davantage de sites de surveillance à long terme sont nécessaires pour mieux évaluer l’évolution de la dynamique des populations d’ombre de fontaine. Les sites de surveillance devraient être situés selon un gradient de pertinence de l’habitat de l’ombre de fontaine, car cela facilitera le repérage rapide des changements au niveau de la répartition.
- Les sites d’échantillonnage devraient être choisis au moyen d’un cadre d’échantillonnage systématique, comme l’échantillonnage aléatoire stratifié (p. ex., classification aquatique).
- La méthode d’échantillonnage utilisée sur chaque site devrait être consignée pour veiller à ce que toutes les données soient collectées au moyen d’une approche uniforme et pour indiquer les cas où l’échantillonnage d’un site se fait par une autre méthode.
- Consigner l’étape du cycle de vie du poisson capturé lors de la prise de mesures individuelles et collectives.
Cela permettra de séparer facilement les données selon l’étape du cycle de vie.

- Prélever des échantillons dans les ruisseaux pendant la saison hivernale pour fournir de l’information sur la répartition hivernale de l’omble de fontaine dans le système.

- Chaque barrage et chaque obstacle existant dans le bassin hydrologique doivent être cartographiés et le degré de connectivité lié à chaque obstacle doit être évalué.

- Les modèles de répartition pourraient être associés à des modèles de décharge d’eaux souterraines; de température des cours d’eau afin de mieux prédire les aires d’occupation.

- Les prédictions des modèles de répartition doivent être validées sur le terrain pour bien évaluer leur efficacité.

**Acknowledgements**

We would like to thank Rob Wilson of the Lake Simcoe Regional Conservation Authority, Brian Morrison of the Ganaraska Region Conservation Authority, Phil Bird of the Credit Valley Conservation Authority, Adam Challice, Gabrielle Gilchrist, and Elizabeth Stanley of the Ministry of Natural Resources and Forestry for providing the electrofishing data and input. Funding for this project was provided by the Lake Simcoe Protection Plan.
## Contents

Summary .......................................................... i
Résumé .......................................................... ii
Acknowledgements ................................................ iii
Introduction ....................................................... 1
  Methods ......................................................... 1
  Habitat Data .................................................. 1
  Brook trout distribution data ................................ 2
  Model selection ............................................... 4
  Model evaluation and calibration ............................ 5
  Model prediction thresholds ................................. 6
  Predicted probability and habitat suitability ............. 7
  Importance of environmental variables .................... 7
  Hindcasting brook trout distribution ....................... 7
  Limitations .................................................... 7

Results .......................................................... 8
  Brook trout distribution data ............................... 8
  Model selection ............................................... 8
  Model performance .......................................... 8
  Hindcasted brook trout distribution ....................... 37

Discussion ....................................................... 44
  The distribution models .................................... 44
  Predicted brook trout distribution ......................... 44
  Management implications ..................................... 46
  Advancing the distribution models ......................... 46

Recommendations ................................................ 51
References ....................................................... 52
Software ........................................................ 58
Introduction

Brook trout (Salvelinus fontinalis) are a coldwater fish species native to eastern North America and have historically been found throughout Ontario (MacCrimmon and Campbell 1969). They are a sought after recreational fish and serve important ecological roles in aquatic ecosystems. However, brook trout populations have been declining or becoming locally extirpated throughout much of their native range. This is largely due to anthropogenic factors, such as land use change, habitat alteration and destruction, fragmentation, invasive species, non-native salmonids, over exploitation and pollution (Hudy et al. 2008). Ontario is no exception and brook trout populations are experiencing increased pressures due to environmental change.

The Lake Simcoe watershed is an area that has undergone many environmental changes due to increasing human development. The changes to the watershed have negatively impacted the fish community and recreational fishery. The environmental degradation in the watershed prompted the creation of the Lake Simcoe Environmental Management Strategy and Lake Simcoe Protection Act (LSPA) in 2008 to guide the development, management and rehabilitation of the watershed. As part of the LSPA, the Lake Simcoe Fish Community Objectives were established to direct the fisheries management decisions of the Ministry of Natural Resources and Forestry (MNRF) as well as other agencies involved with fisheries monitoring and management in the Lake Simcoe watershed. The fish community objectives recognize the vulnerability of brook trout and specify the need to maintain existing brook trout populations and restore coldwater tributary habitat (OMNR 2011). However, there is limited information on the current distribution of brook trout in the Lake Simcoe watershed and this information is required to inform management, conservation, restoration and monitoring efforts.

Species distribution models (SDMs) have many practical applications for conservation and resource management. For example, SDMs have been used to assess the habitat requirements of a species (Buisson et al. 2008), predict the occurrence of rare or endangered species (Guisan et al. 2006), inform conservation and restoration efforts (Estrada et al. 2011; Oppel et al. 2012) and predict the impact of future climate change (Wenger et al. 2011). SDMs involve relating the occurrence of a species (presence/absence) to abiotic and biotic environmental variables. A variety of statistical approaches can be used to model species-habitat relationships and the choice of a particular statistical model will depend on the data available and type of relationships being modelled (Olden and Jackson 2002; Elith and Leathwick 2009). SDMs are a valuable tool for resource managers because they offer a cost effective way to estimate the distribution of a species without having to do an exhaustive habitat assessment of a landscape or watershed.

The purpose of this study is to predict the distribution of brook trout in the Lake Simcoe watershed and to determine which environmental variables influence their occurrence. We did this by fitting several types of SDMs to brook trout occurrence and habitat data from the watershed. These SDMs were then compared to determine which one(s) provided the best predictive performance. The selected model(s) were used to predict the occurrence of brook trout in stream reaches throughout the watershed. We also hindcasted the past distribution of brook trout using the SDMs by artificially maximizing the amount of forest in the watershed (Baker et al. 2005) in an effort to provide context for the current distribution. Knowledge of the distribution and habitat use of brook trout in the Lake Simcoe watershed will help to inform brook trout conservation, restoration and monitoring efforts in the watershed.

Methods

Habitat Data

All habitat data was obtained from the Ontario Aquatic Ecosystem Classification (AEC) system developed by Jones and Schmidt (2016). Catchments and streams were delineated from a digital elevation model with a 30 m resolution using ArcHydro (ArcGIS 10.1; ESRI, 2012). Delineated streams were constrained to those draining a minimum area of 1 km². The AEC system also contains a variety of environmental variables derived from remotely sensed satellite data, existing monitoring networks, or using predictive methods (i.e., modelling or interpolation; Jones and Schmidt 2016). The environmental variables represent geology, hydrology, climate and land use at four different spatial scales: A) upstream catchment area; B) reach contributing area; C) upstream channel area; and D) reach channel area (see illustration). The upstream catchment area encompasses the entire drainage area above the most downstream section of a reach.
The reach contributing area consists of the area that directly drains into the reach of interest. The upstream channel area includes all stream reaches and riparian areas upstream of a reach as represented by a 30 m raster streamline. The reach channel is a confluence-to-confluence unit and its associated riparian buffer (for more detail see Jones and Schmidt 2016).

Prior to using the habitat dataset for any analyses, we first checked the variables for collinearity using a correlation matrix. Any variables with a correlation > |0.7| were extracted from the correlation matrix for further examination (McKenna and Johnson 2011; Aguirre-Gutiérrez et al. 2013). One variable from each extracted pair was retained with preference given to a variable from the highest spatial scale unless an ecological reason was evident to keep a variable from a lower scale (McKenna and Johnson 2011, Kanno et al. 2015b). The data was then checked for variables with little to no informative variation (i.e., contained primarily zeroes). Variables identified as uninformative were removed from the dataset. The habitat variables retained were then visually checked for normality and transformed if necessary. Finally, the habitat variables were standardized by subtracting each variable by its mean and dividing by its standard deviation (z scores). The removal of collinearity and uninformative variables reduced the number of habitat variables down to 19 (Table 1).

**Brook trout distribution data**

The brook trout distribution data used for modelling was collected by a variety of agencies between 1992–2015, including the Ministry of Natural Resources, Fisheries and Oceans Canada and local conservation authorities. The data is stored in the HabProgs database, which is a relational database containing stream faunal and habitat data from across Ontario. The stream data was collected using the Ontario Stream Assessment Protocol (OSAP; Stanfield 2013). The use of the OSAP ensures that sampling methodologies are consistent across agencies and this allows stream assessment data to be reliably combined for large scale analysis. All stream sampling records used for modelling were obtained via electrofishing.

The electrofishing data was divided into individual and bulk catch records within the HabProgs database. These records were combined to create the brook trout occurrence dataset. A site with brook trout present was assigned a ‘1’ and a site absent of brook trout was assigned a ‘0’. Zeroes were not added to streams that had not been sampled. The final brook trout occurrence dataset needed to have one presence/absence record per sampled reach before it could be used for distribution modelling. As such, the dataset had to be further formatted.

The stream sampling records for the Lake Simcoe watershed spanned a time period from 1992–2015. During this time period, several sites were visited on multiple occasions. In these cases, sites were considered to support brook
trout if a brook trout was caught during at least one sampling event. This rule was applied because electrofishing has imperfect detection and an absent designation does not guarantee that brook trout are truly absent (Hense et al. 2010; Mackenzie et al. 2002). In contrast, a positive detection indicates that at least the brook trout caught during a sampling event were occupying the site. In essence, sites which contained only zeros are confounded between true absence of brook trout from the site and false absences where brook trout were present but went undetected. Thus our predictive models based on habitat where brook trout were detected, combined with the knowledge that detection was imperfect, provides a conservative estimate of potential brook trout distribution in the Lake Simcoe watershed.

**Table 1.** The environmental variables used during the development of the brook trout distribution models. The environmental variables for each sample site were derived from one of four spatial scales: the upper catchment area (UCA), upper channel (UCh), reach contributing area (RCA) and reach channel (RCh). Variables with SD indicate standard deviations.

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Scale</th>
<th>Variable Description</th>
<th>Abbreviation</th>
<th>Units</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topology</td>
<td>RCA</td>
<td>Northern Aspect</td>
<td>RAN</td>
<td>% Area</td>
<td>Ln(x+0.01)</td>
</tr>
<tr>
<td></td>
<td>RCA</td>
<td>Southern Aspect</td>
<td>RAS</td>
<td>% Area</td>
<td>Ln(x+0.01)</td>
</tr>
<tr>
<td></td>
<td>RCA</td>
<td>Mean Elevation</td>
<td>ELV</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RCh</td>
<td>Elevation SD</td>
<td>ELV_SD</td>
<td>M</td>
<td>Ln(x+0.01)</td>
</tr>
<tr>
<td></td>
<td>UCA</td>
<td>Max Elevation</td>
<td>ELV_MAX</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geology</td>
<td>Mean Overburden Thickness¹</td>
<td>OVBR</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RCA</td>
<td>Base Flow Index²</td>
<td>BFI</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UCA</td>
<td>Topographic Wetness Index³</td>
<td>TWI</td>
<td>Ln(x+0.01)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UCh</td>
<td>Mean Stream Power Index³</td>
<td>USPI</td>
<td>Ln(x+0.01)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UCh</td>
<td>Mean Channel Slope</td>
<td>CSL</td>
<td>%</td>
<td>Ln(x+0.01)</td>
</tr>
<tr>
<td></td>
<td>RCA</td>
<td>Slope SD</td>
<td>SLP_SD</td>
<td>SQRT</td>
<td></td>
</tr>
<tr>
<td>Climactic</td>
<td>UCA</td>
<td>Air Temperature Range</td>
<td>ATR</td>
<td>°C</td>
<td>SQRT</td>
</tr>
<tr>
<td></td>
<td>UCA</td>
<td>Max Annual Air Temperature</td>
<td>TMAX</td>
<td>°C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UCA</td>
<td>Precipitation Seasonality</td>
<td>PS</td>
<td>Ln(x)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UCA</td>
<td>Max Annual Precipitation</td>
<td>MP</td>
<td>Mm</td>
<td></td>
</tr>
<tr>
<td>Land Use</td>
<td>UCA</td>
<td>Agriculture</td>
<td>AGRI</td>
<td>% Area</td>
<td>Ln(x+0.01)</td>
</tr>
<tr>
<td></td>
<td>UCA</td>
<td>Community Infrastructure</td>
<td>CMIF</td>
<td>% Area</td>
<td>Log₁₀(x+0.01)</td>
</tr>
<tr>
<td></td>
<td>RCA</td>
<td>Swamp</td>
<td>SWP</td>
<td>% Area</td>
<td>Log₁₀(x+0.01)</td>
</tr>
<tr>
<td></td>
<td>RCA</td>
<td>Treed</td>
<td>TREE</td>
<td>% Area</td>
<td>Log₁₀(x+0.01)</td>
</tr>
</tbody>
</table>

¹Mean Overburden Thickness: the depth of glacial materials overlaying bedrock (e.g., kames and moraines).
²Neff et al. (2005).
³Wilson and Gallant (2000).

Air Temperature Range: (Max of mean monthly maximum air temperature) – (minimum of mean monthly minimum air temperature).

Precipitation Seasonality: Coefficient of variation for mean monthly precipitation.

Treed: The combined area of sparse, upland, deciduous, mixed and coniferous treed areas.

SQRT: Square root transformation.

There were also many instances where there were several sites located in a reach. In this scenario, the dataset was pruned so that only one site represented a given reach. The site selected to represent a reach was chosen based on two rules: 1) if sites within a reach show the same brook trout occupancy, then a site was randomly selected to represent the reach; and 2) if sites within a reach show different brook trout occupancy (i.e., present and absent), then a site that shows brook trout as present was selected. A site with brook trout present was chosen to represent a reach based on the same reasoning provided for multiple sampling events per site.
Not all sampling sites were located on a delineated stream reach. Sites located within 80 m of a stream reach were snapped to that reach (most were within 20 m). Sites outside of the 80 m area were identified and snapped to the nearest stream reach according to two criteria: 1) the site must be located within the reach contributing area of the closest stream reach; and 2) the site cannot be snapped from a headwater area to a stream reach of higher order (i.e., cannot be snapped to a stream reach that is not located in a headwater catchment). If any one of these criteria were not satisfied, the sampling site was excluded from the dataset (10 sample sites). These criteria were used to prevent sites from being snapped to stream reaches of unequal stream order that may not have habitat variables consistent with the original site location. A total of six sites were manually snapped to stream reaches using these criteria. Most of these sites were located directly upstream (~100–200 m) of the stream reach they were snapped to.

The formatting and pruning of the brook trout occurrence dataset significantly reduced the number of brook trout records. The original dataset contained 1256 records from 493 different sites over the 24 years of data collection. The final pruned dataset contained 345 records with 72 reaches containing brook trout and 273 reaches without brook trout (Figure 1). The final dataset was then checked for spatial autocorrelation using a variogram to determine if spatial autocorrelation needed to be incorporated into the distribution models.

**Model selection**

The brook trout occurrence data were fitted to different models: logistic regression, general additive, boosted regression tree, random forest and support vector machine models. The predictions from these models were also incorporated into an ensemble prediction. Multiple models were developed in order to identify the model that provided the most robust prediction of brook trout distribution in the Lake Simcoe watershed. All statistical analyses were done using the R Statistical Computing Environment (version 3.2.3; R Core Team 2015).

Logistic regression is one of the most straightforward and widely used techniques for species distribution modelling. The selection of a logistic regression model was done by first fitting the model with all of the habitat predictor variables. The predictor variables were then iteratively removed using a hypothesis testing approach, whereby the predictor variable with the highest p-value according to an analysis of deviance was removed until all predictor variables were statistically significant ($\alpha < 0.05$; Zuur et al. 2009). The final model was then checked for evidence of over-dispersion by ensuring the ratio of residual deviance to degrees of freedom was $\sim 1$ (Zuur et al. 2009). We used a complimentary log-log link function for the logistic regression model because this link function has been found to perform well on skewed response data (Zuur et al. 2009). We also included the best one-variable logistic regression model to use as the minimum performance standard for the other distribution models.

General additive models (GAMs) are often used for species distribution modelling because they are capable of modelling complex non-linear relationships. For model selection, we could not create a full model with all variables smoothed as a non-linear function because this completely over-fitted the data and provided no meaningful results. Instead, we first created four models with each having an equal amount of smoothed variables (~5 per model) and unsmoothed variables (~14 per model). We then fit these four models to determine which variables we should include in the full model as non-linear. This process identified five variables that were included in the full model as non-linear variables and fourteen variables as linear. Model selection began by first evaluating the non-linear variables and removing any with a degree of freedom $\sim 1$ (essentially linear). Model selection then continued using the hypothesis testing approach described above for the logistic regression (Zuur et al. 2009). The GAMs were fit and assessed using the mgcv package in R (Wood 2016).

Boosted regression tree models (BRTs) are a combination of regression and machine learning techniques. The algorithm works by producing many relatively simple classification trees and then combines them to optimize predictive performance (Elith et al. 2008). BRTs have many advantages over other modelling techniques because they are insensitive to the scale of measurement, missing data, outliers and non-normal distributions (Elith et al. 2008). BRTs are also capable of modelling higher order interactions among predictor variables. Several parameters of BRT models have to be optimized during the model selection process, including the learning rate (shrinkage), tree complexity and the bag fraction (see Elith et al. 2008). These parameters are optimized by manually manipulating them until a model reaches peak performance. A BRT model was selected by first fitting a model with all the habitat predictor variables. The
relative importance of predictor variables was assessed based on a measure of the number times a variable was used for splitting, weighted by the squared model improvement due to each split and averaged over all trees (Friedman and Meulman 2003). Variables that contributed less than 5% to the full model were removed (Buston and Elith 2011). The simplified BRT was then fit and the model parameters optimized. The simplified model was kept if its performance was equal to or greater than the full model as measured by the deviance explained. The BRT models were fit and assessed using the gbm and dismo packages in R (Ridgeway 2015; Hijmans et al. 2016).

Random forests are a classification and regression tree technique that “grows” many classification trees and combines them to create “forest”. Each classification tree is built using a bootstrapped sample of the dataset and a random set of predictor variables (Breiman 2001). The trees are grown to maximum size without pruning and then aggregated by averaging the trees (Breiman 2001). The random predictor selection reduces bias during model fitting and prevents the algorithm from overfitting the data (Breiman 2001; Prasad et al. 2006). A full random forest model was fitted to the brook trout data and the number of trees and the number of predictor variables used per split was optimized using the out-of-bag error rates. Variable importance for a random forest model is assessed by the reduction in model performance when a variable is permuted (Prasad et al. 2006). Variable importance values from the full model were used to build several simplified candidate models, where a simplified model was built every time there was a distinct drop in the ordered variable importance values (Genuer et al. 2010). This resulted in four candidate models with 5-12 variables. The performance of the simplified candidate models and the full model were then compared using the out-of-bag error rates and the model with the lowest error rate was selected (Genuer et al. 2010). The random forest models were fit and assessed using the party package in R (Hothorn et al. 2015).

Support vector machines (SVMs) have become a more popular means of modelling species distributions and are another form of machine learning (Drake et al. 2006). SVMs use a kernel to map many variables into a hyperspace where complex patterns are more simply represented (Drake et al. 2006). A full SVM model was fit using all the habitat variables. The cost and gamma parameters of the full model were optimized by finding the combination that provided the lowest classification error (see the tune function; Meyer et al. 2015). The importance of each variable in the full model was then ranked using a procedure proposed by Guyon et al. (2002). Variables were then removed from the model starting with the least important variables until the performance of the reduced model fell below the full model. The performance of the SVM models was evaluated using the area under the receiver operating characteristic curve (AUC; see Model Evaluation and Calibration). The SVMs were fit and tuned using the e1071 package in R (Meyer et al. 2015).

We also combined predictions from all models into an ensemble prediction because such an approach has been shown to provide more accurate predictions (Araujo and New 2007; Marmion et al. 2009; Thuiller et al. 2009; Oppel et al. 2012). A predictive model provides a simplistic representation of a complex biological system and each model can be seen as one possible state of that system (Araujo and New 2007). Ensemble predictions are often more accurate because they incorporate many possible system states. The ensemble predictions were created using a weighted average consensus method (Marmion et al. 2009). The predictions from each model were weighted by the discriminatory power of the model as measured by the AUC (Marmion et al. 2009; Oppel et al. 2012). The AUC weights were squared to put more emphasis on predictions from models with better discriminatory power.

Model evaluation and calibration

The models were evaluated using a variety of metrics that assessed the calibration and discriminatory ability of each model. The calibration of a model is how well the predicted probability of occurrence represents the proportion of sites occupied (Pearce and Ferrier 2000; Phillips and Elith 2010; Jiménez-Valverde et al. 2013). In other words, the calibration can be considered as the reliability of the prediction in relation to the actual proportion of species occurrence. The discriminatory ability is a model’s ability to correctly classify a site as occupied or unoccupied (Pearce and Ferrier 2000). The metrics used to assess model discrimination were sensitivity, specificity, Cohen’s kappa, the true skill statistic and the AUC. The sensitivity and specificity are a measure of the correct positive predictions and correct negative predictions produced by the models, respectively. The kappa statistic is a chance-corrected measure of model accuracy that ranges between -1 and 1, where values ≤ 0 indicate a model that performs no better than chance and a value of +1 indicates a model with perfect model prediction (Cohen 1960, Allouche et al. 2006). Kappa statistic values between 0.4 and 0.75 are considered a good model fit and values >0.75 are considered an excellent model fit (see Fielding and Bell
1997). However, the kappa statistic has been criticized because it is sensitive to the prevalence (the overall proportion of locations where the species is predicted to be present) of the modeled species (McPherson et al. 2004; Allouche et al. 2006; McPherson and Jetz 2007). The true skill statistic was developed to address the bias of the kappa statistic and ranges from -1 to 1, where values ≤ 0 indicate a model that performs no better than chance and +1 indicates a model with perfect model predictions (Allouche et al. 2006). The same model performance classification scale used for the kappa statistic also applies to the true skill statistic. The AUC is one of the most popular model accuracy metrics. Values range from 0.5-1.0; 0.5 indicates a model that performs no better than chance and a value of 1 indicates a model with perfect prediction accuracy (Wenger et al. 2011; Wagner et al. 2013; DeWeber and Wagner 2015; Kanno et al. 2015b). Models with AUC values between 0.7 and 0.9 are considered to have good discrimination abilities and models with AUC values > 0.90 are considered to have excellent discrimination abilities (Pearce and Ferrier 2000). Unlike the other discrimination metrics, the AUC value is independent of the probability threshold chosen for classification (Fielding and Bell 1997; Pearce and Ferrier 2000). All metrics of discrimination were calculated using the PresenceAbsence package in R (Freeman and Moisen, 2012).

Model calibration metrics have been used much less frequently than discrimination metrics to assess species distribution models. However, model calibration metrics can identify important weaknesses of distribution models (Pearce and Ferrier 2000; Phillips and Elith 2010). The calibration of the brook trout distribution models was evaluated using several metrics: bias, spread and point-biserial correlation. To assess model calibration, model prediction probabilities are organized into bins and the mean probability value of each bin is plotted against the fraction of true presences (Pearce and Ferrier 2000). A perfectly calibrated model will have all the data fall on a 45° line representing a slope of 1 and an intercept of 0. A linear model is fitted to the binned data and the slope and intercept of the linear model provides the spread and bias calibration measures, respectively. A slope greater than 1 indicates a model that overestimates predictions above 0.5 and underestimates predictions less 0.5 (vice versa for slope < 1; Pearce and Ferrier 2000). The measure of bias describes the consistent overestimation or underestimation of predicted values, where a positive bias indicates overestimation and a negative bias indicates underestimation (Pearce and Ferrier 2000). A point-biserial correlation between the presence-absence data and the model predictions is a measure of accuracy that takes into account the deviation of predicted values from observed values (Elith 2006; Phillips and Elith 2010).

Model validation was done using a k-fold cross-validation procedure. Often, a test dataset is withheld that is never used during the model training process. However, our brook trout distribution dataset was relatively small (Figure 1) and the lack of brook trout occurrences prevented the use of an exclusive test dataset. The k-fold cross-validation procedure provides an efficient alternative to validate predictive models while also testing the model on all available data (Arlot and Celisse 2010). A 5-fold cross-validation was used in this study. The procedure involves splitting the brook trout dataset into 5 folds that are stratified by brook trout presence. Stratification ensures that the proportion of brook trout occurrence in each fold is consistent with the original dataset (~21% of observations). Each fold is then used as a test dataset while the remaining folds are combined to form a training dataset (5 iterations total). The discrimination metrics are calculated for each test set (fold) during the procedure. The output can then be used to calculate the mean model performance and its uncertainty. The calibration metrics were not included in the cross-validation procedure because there was not enough presence data in each fold to accurately estimate the calibration metrics.

**Model prediction thresholds**

The conversion of predicted probabilities to a binary presence/absence designation requires the application of a probability threshold. The choice of a probability threshold depends on the conservation question being asked. If you are interested in identifying stream reaches for brook trout reintroduction and need to ensure areas identified possess suitable habitat, then a higher probability threshold should be selected to restrict sites to those with only the best habitat (Pearce and Ferrier 2000). However, if you would like to identify sites of potential brook trout occupation, then the choice of a lower probability threshold would be advised because you would select stream reaches spanning the full range of habitat suitability (Pearce and Ferrier 2000). We chose to evaluate several probability thresholds to determine how they impacted model performance and to help us determine what threshold is most appropriate. When a species’ prevalence is low, thresholds that make the predicted prevalence equal to the observed prevalence or maximize the kappa statistic have been shown to perform well (Freeman and Moisen 2008). A threshold that minimizes the distance to the top left corner of the receiver operating characteristic curve has also been shown to perform well \((1-\text{sensitivity})^2 + (1-\text{specificity})^2\);
Liu et al. 2005). We used the model performance metrics described above to evaluate the impact the three threshold criteria had on model performance.

**Predicted probability and habitat suitability**

The use of a binary presence/absence designation does not necessarily provide any information about the habitat suitability of a reach for brook trout. This type of information would be useful to help prioritize conservation efforts in the Lake Simcoe watershed. The predicted probability of brook trout occurrence can be used as a rough measure of habitat suitability. A reach with a high probability of brook trout presence can be considered more suitable than a reach with a low probability. However, the probability values should be interpreted in relation to the probability threshold used to convert the probabilities into presence/absence data (Jimenez-Valverde and Lobo 2006). If a probability threshold is < 0.5 then sites that are potentially suitable for brook trout will also have predicted probabilities < 0.5.

**Importance of environmental variables**

We were also interested in determining how important the habitat variables were within and among the distribution models because this will help indicate which variables are the primary factors determining brook trout occupancy. Each distribution model was trained to the dataset and a standard prediction was made. A second prediction was then made using the dataset after randomizing one of the habitat variables (Thuiller et al. 2010). The standard prediction was then correlated with the new prediction. The correlation coefficient provides an indication of variable importance (i.e., a low correlation between predictions indicates an important habitat variable; Thuiller et al. 2010). The correlation coefficients were subtracted from one, which gave variables with the most importance the highest values. Habitat variable importance was not calculated for the ensemble model because there is no meaningful way to make the calculation (Aguirre-Gutiérrez et al. 2013).

**Hindcasting brook trout distribution**

Hindcast distribution modelling was used to predict what the brook trout distribution would have been prior to human development (Baker et al. 2005; Kilgour and Stanfield 2006; Maire et al. 2015). The proportion of forested landscape for all reach contributing areas in the Lake Simcoe watershed was increased to one minus the proportion of swamp (i.e., all land not swamp is forested). The logistic regression and support vector machine include the amount of community infrastructure and agriculture as landscape variables. For these models, the community infrastructure and agriculture was set to zero (i.e., no human development). The distribution of brook trout in the Lake Simcoe watershed has changed with continued development and artificially changing the watershed back to forested land will simulate what the maximum potential brook trout distribution may have been before human development.

**Limitations**

The brook trout occurrence data was not specifically collected for the purpose of brook trout distribution modelling. Instead, sampling sites were primarily chosen based on convenience sampling or as part of a specific project and not according to a systematic sampling methodology (i.e., stratified random sampling). As such, we acknowledge that there may be inherent biases in the data and the models may be less informative than a model based on a more robust survey design. However, these models were built using the best available data and will still be highly informative for conservation purposes.

There was also data on the effort required to capture individuals for many of the sites in the watershed. However, there were errors and inconsistencies in the data, which led to a reduced dataset of 70 sites. This was not enough data to build a reliable predictive model for brook trout catch per unit effort (CPUE). We mapped the CPUE data to provide a visual of brook trout abundance in the Lake Simcoe watershed.

The distribution models do not discriminate among brook trout life-stages because there was not a sufficient amount of data about the life-stage of captured individuals. There were some occurrences where brook trout were identified as young-of-the-year (YOY). The YOY tend not to travel long distances from their hatching area and the presence of
YOY can be used to locate potential spawning areas (Petty et al. 2005). We mapped the YOY records to provide an indication of where spawning sites might be located in the Lake Simcoe watershed.

**Results**

**Brook trout distribution data**

The brook trout distribution data used for modelling had good coverage across the majority of the Lake Simcoe watershed with sampling reaches located in every major stream system (45 systems; Figure 1 and 2). The number of sampling reaches located in each stream system varied from 1–48 with the largest stream systems receiving the most sample sites (i.e., the Beaver, Black, West Holland, East Holland and Pefferlaw Rivers; Figure 2). However, most stream systems were relatively under sampled and had <10 sample reaches (78%; Figure 2). The number of times a reach was visited during 1992–2015 varied among sites with 60% of the sites having a single sampling event (Figure 3).

The prevalence of brook trout in the Lake Simcoe watershed is low. Of the 345 reaches in the brook trout distribution dataset, only 72 had brook trout present during at least one sampling event. The sampling reaches with brook trout present are located in a limited number of stream systems (36% of stream systems; Figure 4) and they are generally limited to headwater reaches (Figure 1). The distribution of brook trout in the Lake Simcoe watershed does not have any strong spatial autocorrelation (Figure 5). As such, we did not explicitly incorporate spatial autocorrelation when modelling brook trout distribution.

**Model selection**

The number of habitat variables included in the final distribution models varied considerably (Table 2). For instance, the general additive model included seven habitat variables, whereas the support vector machine incorporated all the habitat variables. The most consistent variables selected among the distribution models were the overburden thickness and stream power index, which were the only variables selected for all models. Other habitat variables that were consistently selected were maximum catchment elevation, maximum monthly air temperature, precipitation seasonality, channel slope, base-flow index and the treed area (selected for four of five models).

**Model performance**

*Threshold Independent Measure*

The AUC of the distribution models during training were high (range 0.89–0.99) and all models were greater than the one-variable logistic regression model (0.78; Figure 6). The AUC values of the distribution models during cross-validation were lower relative to when the models were fit using the entire training dataset (range 0.86–0.89; Figure 6). However, the cross-validation AUC values were still high and indicate that the distribution models perform well on withheld data. The cross-validated AUC values of the distribution models did not differ among models (Figure 6; pairwise z-tests: P >0.05). The reduction in the AUC values between training and cross-validation indicates that there was some evidence of model over-fitting, but this over-fitting has not diminished the predictive capacity of the models to any great extent (AUC values remained high).

*Threshold Dependent Measures*

Unlike the AUC, the sensitivity, specificity, kappa statistic and true skill statistic (TSS) are all threshold dependent measures of model performance. The performance of the individual distribution models was consistent within each of the threshold criteria regardless of the performance metric considered (Figure 7-9). There were no significant differences among the distribution models when comparing the cross-validated performance metrics within each threshold criteria (pairwise z-tests: P >0.05). In general, there was more variation among threshold criteria than among individual distribution models within a threshold criterion. Because the individual distribution models were comparable within threshold criteria, we focused on comparing the general performance of the three threshold criteria.
Figure 1. Brook trout distribution in the Lake Simcoe watershed. Each point represents the presence (blue) or absence (green) of brook trout among the sampled reaches.
Figure 2. Number of electrofishing sample sites located in each stream after dataset pruning.
Figure 3. Histogram of electrofishing sampling frequency for all sampling sites in the Lake Simcoe watershed from 1992-2015.
Figure 4. Proportion of sampled reaches within a stream system that have brook trout present.
Figure 5. Variogram of the brook trout distribution data for the Lake Simcoe watershed using geodesic distance (metres).
### Table 2
The environmental variables included in the one-variable logistic regression (LR_Simple), multi-variable logistic regression (LR), general additive (GAM), boosted regression tree (BRT), random forest (RF) and support vector machine (SVM) models. Parameters with ‘s( )’ are smoothed terms from the general additive model. The variables are not listed in any particular order. See Table 1 for variable descriptions.

<table>
<thead>
<tr>
<th>LR_Simple</th>
<th>LR</th>
<th>GAM</th>
<th>BRT</th>
<th>RF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVBR</td>
<td>RAS</td>
<td>OVBR</td>
<td>OVBR</td>
<td>OVBR</td>
<td>OVBR</td>
</tr>
<tr>
<td>ELV_MAX</td>
<td>TREE</td>
<td>TREE</td>
<td>BFI</td>
<td>PS</td>
<td></td>
</tr>
<tr>
<td>OVBR</td>
<td>RAS</td>
<td>BFI</td>
<td>ELV_MAX</td>
<td>ATR</td>
<td></td>
</tr>
<tr>
<td>BFI</td>
<td>USPI</td>
<td>ATR</td>
<td>CSL</td>
<td>BFI</td>
<td></td>
</tr>
<tr>
<td>USPI</td>
<td>s(EVL_SD)</td>
<td>USPI</td>
<td>ATR</td>
<td>TMAX</td>
<td></td>
</tr>
<tr>
<td>CSL</td>
<td>s(SLP_SD)</td>
<td>CSL</td>
<td>TMAX</td>
<td>ELV_MAX</td>
<td></td>
</tr>
<tr>
<td>TAMX</td>
<td>s(PS)</td>
<td>PS</td>
<td>USPI</td>
<td>CSL</td>
<td></td>
</tr>
<tr>
<td>CMIF</td>
<td></td>
<td>SLP_SD</td>
<td>TREE</td>
<td>TWI</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ELV_SD</td>
<td>PS</td>
<td>SLP_SD</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ELV_MAX</td>
<td>ELV_MEAN</td>
<td>CMIF</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TMAX</td>
<td>TWI</td>
<td>MP</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MP</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ELV_MEAN</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TREE</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>USPI</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ELV_SD</td>
<td>AGRI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>RAN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SWP</td>
</tr>
</tbody>
</table>
Figure 6. The area under the receiver operating characteristic curve (AUC) for the one-variable logistic regression (LR_Simple), multi-variable logistic regression (LR), general additive (GAM), boosted regression tree (BRT), support vector machine (SVM), random forest (RF) and ensemble models. The points represent model AUC values that were calculated for each model type using 5-fold cross validation (blue) and the full dataset (red). The error bars on the cross validation statistics represent standard deviations.
Figure 7. A comparison of model performance using a) sensitivity, b) specificity, c) the kappa statistic and d) the true skill statistic among one-variable logistic regression (LR_Simple), multi-variable logistic regression (LR), general additive (GAM), boosted regression tree (BRT), support vector machine (SVM), random forest (RF) and ensemble models. The performance metrics were calculated using a probability threshold that maximized the kappa statistic. The points represent model performance metrics that were calculated for each model type using 5-fold cross validation (blue) and the full dataset (red). The error bars on the cross validation statistics represent standard deviations.
Figure 8. A comparison of model performance using a) sensitivity, b) specificity, c) the kappa statistic and d) the true skill statistic among one-variable logistic regression (LR_Simple), multi-variable logistic regression (LR), general additive (GAM), boosted regression tree (BRT), support vector machine (SVM), random forest (RF) and ensemble models. The performance metrics were calculated using a probability threshold where the predicted prevalence equals the observed prevalence. The points represent model performance metrics that were calculated for each model type using 5-fold cross validation (blue) and the full dataset (red). The error bars on the cross validation statistics represent standard deviations.
Figure 9. A comparison of model performance using a) sensitivity, b) specificity, c) the kappa statistic and d) the true skill statistic among one-variable logistic regression (LR_Simple), multi-variable logistic regression (LR), general additive (GAM), boosted regression tree (BRT), support vector machine (SVM), random forest (RF) and ensemble models. The performance metrics were calculated using a probability threshold that minimizes the distance to the top left corner of the receiver operating characteristic curve \([(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2]\). The points represent model performance metrics that were calculated for each model type using 5-fold cross validation (blue) and the full dataset (red). The error bars on the cross validation statistics represent standard deviations.
The sensitivity of the distribution models was modest with the mean cross-validated sensitivity ranging from 0.55–0.78 (Figure 7–9). The cross-validated sensitivity was highest for the threshold that minimized the distance to the upper left corner of the receiver operating characteristic curve (minROC; range 0.71–0.78), whereas the maximum kappa threshold (MK; range 0.55–0.64) and the predicted prevalence equal to the observed prevalence threshold (PredObs; range 0.58–0.70) had lower sensitivities. The sensitivity of the distribution models when fit with the training dataset was similar to the cross-validated sensitivity.

The specificity of the distribution models was higher than the sensitivity with the mean cross-validated specificity ranging from 0.79–0.94 (Figure 7–9). The MK threshold (range 0.87–0.94) and the PredObs threshold (range 0.88–0.90) had the highest cross-validated specificity, whereas the minROC threshold has the lowest specificity (range 0.79–0.88). The specificity of the distribution models when fit with the training dataset was similar to the cross-validated specificity.

The kappa statistic and the TSS are both measures of model accuracy and these metrics show no distinct patterns among the three thresholds (Figures 7–9). The kappa statistic and the TSS were consistent among the thresholds and all indicated that the distributions models had a fair to good performance (kappa statistic >0.45 and TSS >0.45). Similar to the sensitivity and specificity, the kappa statistic and TSS were similar between the distribution models when fit using the training dataset and when using cross-validation.

Model Calibration

Model calibration varied among the distribution models (Figure 10, Table 3). The one-variable logistic regression model, the multi-variable logistic regression model and the general additive model all had a spread value very close to one. In contrast, the boosted regression model, random forest model, support vector machine and ensemble model had spread values larger than one indicating that predicted probabilities above 0.5 are overestimated and the predicted probabilities below 0.5 are underestimated (Figure 10). The bias of the distribution models was very low and did not reveal any consistent deviation in the predicted probabilities. The point-biserial correlations were more variable among models than the other calibration metrics. The boosted regression tree model had the strongest point-biserial correlation, whereas the one-variable logistic regression model had the weakest (Table 3). There was not a single best performing model across all three calibration metrics. However, the support vector machine had the best overall ranking across the three calibration categories, whereas the random forest and general additive models had the worst overall ranking.
Figure 10. Binned calibration plots of the observed proportion of positive cases versus the predicted probability of occurrence for the a) one-variable logistic regression model, b) multi-variable logistic regression model, c) general additive model, d) boosted regression tree model, e) support vector model, f) random forest model and g) an ensemble model incorporating the results of models b to f.
Table 3. The probability threshold, predicted prevalence and probability range for the logistic regression (LR), general additive (GAM), boosted regression tree (BRT), random forest (RF), support vector machine (SVM) and ensemble (EM) distribution models.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>COR</th>
<th>Spread</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple_LR</td>
<td>0.41</td>
<td>1.03</td>
<td>-0.0052</td>
</tr>
<tr>
<td>LR</td>
<td>0.62</td>
<td>1.02</td>
<td>-0.015</td>
</tr>
<tr>
<td>GAM</td>
<td>0.74</td>
<td>1.06</td>
<td>-0.037</td>
</tr>
<tr>
<td>BRT</td>
<td>0.90</td>
<td>1.28</td>
<td>-0.037</td>
</tr>
<tr>
<td>RF</td>
<td>0.76</td>
<td>1.36</td>
<td>-0.015</td>
</tr>
<tr>
<td>SVM</td>
<td>0.75</td>
<td>1.23</td>
<td>0.00027</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.81</td>
<td>1.26</td>
<td>-0.016</td>
</tr>
</tbody>
</table>

Predicted brook trout distribution

The primary goal for this project was to accurately predict where brook trout are located in the Lake Simcoe watershed. As such, we needed an accurate distribution model with a high sensitivity (true positive rate). We chose the minROC threshold as the probability threshold to predict the distribution of brook trout in the Lake Simcoe watershed because it provided the highest sensitivity (≥0.71), while also having a reasonable specificity (≥0.79).

We predicted the distribution of brook trout using all the individual distribution models (Figure 11–16). The predicted prevalence of brook trout varied among the distribution models with the boosted regression tree model having the lowest predicted prevalence and the support vector machine having the highest predicted prevalence (Table 4). The distribution models agreed on the predicted presence of brook trout in 116 reaches (188 reaches when including known presence sites; Figure 17). The most confidence can be placed on the predicted presence of brook trout where all the distribution models agree and this confidence diminishes with less agreement among models.

We also mapped the predicted probability of brook trout occurrence for each reach to provide a measure of habitat suitability (Figure 18–23). The threshold probabilities of the distribution models were low (0.19–0.30) due to the low prevalence of brook trout in the watershed (Table 4). As a result, the predicted probability of brook trout sites will also extend into the lower range of probability values (<0.50) and this must be considered when interpreting the probability maps.

The distribution of brook trout is limited in the Lake Simcoe watershed and brook trout are primarily located in the headwater areas (Figure 11–17). Notable exceptions are Bluff’s Creek (immediately south of Orillia), Kidd’s Creek (downtown Barrie) and Lovers Creek (south side of Kempenfelt Bay), where the majority of these systems are suitable for brook trout. The headwater region of the Pefferlaw River possessed the most suitable habitat for brook trout with many reaches predicted to support brook trout. The Black River, West Holland River, East Holland River and Uxbridge Brook also had brook trout present in the headwater regions, but to a lesser extent than the Pefferlaw River. The northeastern corner of Lake Simcoe is mostly absent of brook trout with only the support vector machine indicating the potential presence of brook trout habitat. However, the support vector machine probability map indicates that the suitability of that area is marginal (Figure 22). The point-biseral correlation (COR), spread and bias of the one-variable logistic regression (Simple_LR), multi-variable logistic regression (LR), general additive (GAM), boosted regression tree (BRT), support vector machine (SVM), random forest (RF) and ensemble models. In general, brook trout were positively related to overburden thickness, stream power index, maximum catchment elevation, channel slope, base-flow index and the proportion of treed landscape, whereas they were negatively related to the maximum monthly air temperature (Figure 24). The overburden thickness was the most important habitat variable for all the distribution models (Table 5). The stream power index was also an important habitat variable and was consistently within the top four important variables (Table 5). The rest of the aforementioned habitat variables varied considerably in their importance among the distribution models (Table 5).
Figure 11. Predicted brook trout distribution in the Lake Simcoe watershed using a logistic regression model.
Figure 12. Predicted brook trout distribution in the Lake Simcoe watershed using a general additive model.
Figure 13. Predicted brook trout distribution in the Lake Simcoe watershed using a boosted regression tree model.
Figure 14. Predicted brook trout distribution in the Lake Simcoe watershed using a random forest model.
Figure 15. Predicted brook trout distribution in the Lake Simcoe watershed using a support vector machine.
Figure 16. Predicted brook trout distribution in the Lake Simcoe watershed using an ensemble of all model predictions.
Table 4. The probability threshold, predicted prevalence and probability range for the logistic regression (LR), general additive (GAM), boosted regression tree (BRT), random forest (RF), support vector machine (SVM) and ensemble (EM) distribution models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Threshold</th>
<th># Present</th>
<th># Absent</th>
<th>Prevalence (%)</th>
<th>Probability Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.27</td>
<td>368</td>
<td>1544</td>
<td>19.2</td>
<td>0.00-1.00</td>
</tr>
<tr>
<td>GAM</td>
<td>0.21</td>
<td>433</td>
<td>1479</td>
<td>22.6</td>
<td>0.00-1.00</td>
</tr>
<tr>
<td>BRT</td>
<td>0.30</td>
<td>263</td>
<td>1649</td>
<td>13.8</td>
<td>0.00-1.00</td>
</tr>
<tr>
<td>RF</td>
<td>0.20</td>
<td>369</td>
<td>1543</td>
<td>19.3</td>
<td>0.00-0.89</td>
</tr>
<tr>
<td>SVM</td>
<td>0.19</td>
<td>449</td>
<td>1463</td>
<td>23.5</td>
<td>0.00-0.98</td>
</tr>
<tr>
<td>EM</td>
<td>0.21</td>
<td>406</td>
<td>1506</td>
<td>21.2</td>
<td>0.00-0.96</td>
</tr>
</tbody>
</table>

Note: The number of present and absent include the occurrence data used to train the distribution models (Present = 72; Absent = 273).
Figure 17. Model consensus among the brook trout distribution models in the Lake Simcoe watershed. The colours of the reach contributing areas indicate how many distribution models agreed on the predicted presence of brook trout.
Figure 18. Predicted probability of brook trout occurrence for the Lake Simcoe watershed using a logistic regression model.
Figure 19. Predicted probability of brook trout occurrence for the Lake Simcoe watershed using a general additive model.
Figure 20. Predicted probability of brook trout occurrence for the Lake Simcoe watershed using a boosted regression tree model.
Figure 21. Predicted probability of brook trout occurrence in the Lake Simcoe watershed using a random forest model.
Figure 22. Predicted probability of brook trout occurrence for the Lake Simcoe watershed using a support vector machine.
Figure 23. Predicted probability of brook trout occurrence for the Lake Simcoe watershed using an ensemble of all model predictions.
Figure 24. Partial dependence plots from the boosted regression tree model for the overburden thickness (OVBR), treed area (TREE), base-flow index (BFI), channel slope (CSL), precipitation seasonality (PS), slope standard deviation (SLP_SD), elevation standard deviation (ELV_SD), stream power index (USPI), air temperature range (ATR), maximum elevation (ELV_MAX) and maximum monthly air temperature (TMAX). The percentages indicate the relative importance of the habitat variables. Marginal effects are standardized to a mean of zero. The habitat variables were transformed and standardized prior to analysis.
Table 5. Habitat variable importance for the logistic regression (LR), general additive (GAM), boosted regression tree (BRT), random forest (RF) and support vector machine (SVM) models. The variables are ordered by their importance within each of distribution models.

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>GAM</th>
<th>BRT</th>
<th>RF</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVBR</td>
<td>0.553</td>
<td>OVBR</td>
<td>0.539</td>
<td>OVBR</td>
<td>0.082</td>
</tr>
<tr>
<td>TMAX</td>
<td>0.252</td>
<td>PS</td>
<td>0.234</td>
<td>USPI</td>
<td>0.079</td>
</tr>
<tr>
<td>USPI</td>
<td>0.248</td>
<td>SLP_SD</td>
<td>0.233</td>
<td>TREE</td>
<td>0.062</td>
</tr>
<tr>
<td>CSL</td>
<td>0.137</td>
<td>USPI</td>
<td>0.109</td>
<td>PS</td>
<td>0.047</td>
</tr>
<tr>
<td>ELV_MAX</td>
<td>0.090</td>
<td>RAS</td>
<td>0.086</td>
<td>BFI</td>
<td>0.046</td>
</tr>
<tr>
<td>BFI</td>
<td>0.076</td>
<td>TREE</td>
<td>0.080</td>
<td>EVL_SD</td>
<td>0.046</td>
</tr>
<tr>
<td>RAS</td>
<td>0.054</td>
<td>EVL_SD</td>
<td>0.063</td>
<td>CSL</td>
<td>0.043</td>
</tr>
<tr>
<td>CMIF</td>
<td>0.033</td>
<td>SLP_SD</td>
<td>0.036</td>
<td>TMAX</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td>ELV_MAX</td>
<td>0.014</td>
<td>MP</td>
<td>0.004</td>
<td>RAN</td>
</tr>
<tr>
<td></td>
<td>ATR</td>
<td>0.007</td>
<td>ATR</td>
<td>0.003</td>
<td>MP</td>
</tr>
<tr>
<td></td>
<td>TWI</td>
<td>0.002</td>
<td>PS</td>
<td>0.008</td>
<td>SWP</td>
</tr>
<tr>
<td></td>
<td>RAS</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CMIF</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SLP_SD</td>
<td>0.009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TWI</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TREE</td>
<td>0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGRI</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATR</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>0.553</td>
<td>0.539</td>
<td>0.082</td>
<td>0.142</td>
<td>0.381</td>
</tr>
<tr>
<td>Min</td>
<td>0.033</td>
<td>0.063</td>
<td>0.007</td>
<td>0.002</td>
<td>0.003</td>
</tr>
</tbody>
</table>

**Hindcasted brook trout distribution**

The hindcasted brook trout distribution for the Lake Simcoe watershed was greater than the current brook trout distribution (Figure 25–30). The mean predicted prevalence of brook trout decreased substantially from 38.4% of all reaches for the hindcasted distribution models to 19.9% for the current distribution models (Table 6). There was a mean habitat loss of 46.8% between the hindcasted and current distribution models (Table 6). The hindcasted distribution of brook trout expanded in the western section of the Lake Simcoe watershed to include most reaches in the area of Lover’s Creek, Kidd’s Creek, Hawkestone Creek and Bluff’s Creek. The southern section of the Lake Simcoe watershed had brook trout expand in the headwater regions of the West Holland, East Holland and Black Rivers. The Pefferlaw River showed the greatest increase in brook trout distribution with the expansion of brook trout from the headwater regions all the way down to the shores of Lake Simcoe. The distribution of brook trout did not make any significant changes in the northeastern section of the watershed and brook trout continued to be absent from this area.
Figure 25. The hindcasted brook trout distribution in the Lake Simcoe watershed using the logistic regression model.
Figure 26. The hindcasted brook trout distribution in the Lake Simcoe watershed using the general additive model.
Figure 27. The hindcasted brook trout distribution in the Lake Simcoe watershed using the boosted regression tree model.
Figure 28. The hindcasted brook trout distribution in the Lake Simcoe watershed using the random forest model.
Figure 29. The hindcasted brook trout distribution in the Lake Simcoe watershed using the support vector machine.
Figure 30. The hindcasted brook trout distribution in the Lake Simcoe watershed using an ensemble of the model predictions.
Table 6. The number of reaches with brook trout present/absent and the predicted prevalence of brook trout for the current and
hindcasted logistic regression (LR), general additive (GAM), boosted regression (BRT), random forest (RF), support vector
machine (SVM) and ensemble (EM) distribution models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Current Present</th>
<th>Current Absent</th>
<th>Current Prev. (%)</th>
<th>Hindcast Present</th>
<th>Hindcast Absent</th>
<th>Hindcast Prev. (%)</th>
<th>Reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>368</td>
<td>1544</td>
<td>19.2</td>
<td>952</td>
<td>49.8</td>
<td>61.3</td>
<td></td>
</tr>
<tr>
<td>GAM</td>
<td>433</td>
<td>1479</td>
<td>22.6</td>
<td>684</td>
<td>35.8</td>
<td>36.7</td>
<td></td>
</tr>
<tr>
<td>BRT</td>
<td>263</td>
<td>1649</td>
<td>13.8</td>
<td>609</td>
<td>31.9</td>
<td>56.8</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>369</td>
<td>1543</td>
<td>19.3</td>
<td>552</td>
<td>28.9</td>
<td>33.2</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>449</td>
<td>1463</td>
<td>23.5</td>
<td>760</td>
<td>39.7</td>
<td>40.9</td>
<td></td>
</tr>
<tr>
<td>EM</td>
<td>406</td>
<td>1506</td>
<td>21.2</td>
<td>841</td>
<td>44.0</td>
<td>51.7</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>381</td>
<td>1531</td>
<td>19.9</td>
<td>733</td>
<td>38.4</td>
<td>46.8</td>
<td></td>
</tr>
</tbody>
</table>

Discussion

The distribution models

The distribution models had strong discrimination abilities regardless of model type (AUC > 0.8, Kappa > 0.45, TSS
> 0.45) and the performance of the distribution models was equal to or better than previous models predicting the
distribution of brook trout (Al-Chokhachy et al. 2013; Stoneman and Jones 2000; Kocovsky and Carlisle 2006; Stanfield
et al. 2006; Steen et al. 2006; Wanger et al. 2011; Wagner et al. 2013; DeWeber and Wagner 2015; Kanno et al.
2015b; Hudy et al. 2008). The sensitivity of the distribution models was lower than the specificity, but this was expected
given that the brook trout dataset contained many more brook trout absences than presences (Brenner et al. 1997;
McPherson et al. 2004). The good performance of the distribution models in this study could be related to the limited
spatial extent used for predicting brook trout distribution (Hernandez et al. 2006). However, the impact geographic extent
has on the accuracy of distribution models is a topic of continued debate (Raes and ter Steege 2007).

The distribution models were also well calibrated and there was no consistent bias in the predicted probabilities. The
machine learning models did have spread values greater than one indicating that they overestimated probabilities greater
than 0.5 and underestimated probabilities less than 0.5. The higher spread values of the machine learning techniques
could result in models that slightly underestimate the predicted distribution of brook trout because the probabilities that
are underestimated are near the threshold probability. However, the departure of the spread values from one is small and
we are confident that the machine learning models produced valid predictions.

Predicted brook trout distribution

The distribution of brook trout was primarily restricted to the headwater areas of the Lake Simcoe watershed, which is
consistent with other regions in the species native range (Stanfield et al. 2006; McKenna and Johnson 2011; Kanno et
al. 2015b). Stanfield et al. (2006) predicted the distribution of brook trout on the south side of the Oak Ridges Moraine
and found that brook trout were mostly located in headwater regions. These headwater regions also possessed
the highest density of brook trout (Stanfield et al. 2006). Headwater regions are generally associated with areas of
groundwater discharge and intact riparian vegetation, which help to mediate summer temperature increases and
provide suitable habitat for brook trout (Power et al. 1999; Gomi et al. 2002; Gerber and Howard 2002; Winter 2007).
In addition, these headwater streams support brook trout spawning habitat (Witzel and MacCrinnon 1983; Curry et al.
2002; Curry and Noakes 1995), which is essential for the maintenance of self-sustaining brook trout populations.

Brook trout occurrence was strongly related to overburden thickness and base-flow index (BFI), which reflects the
association between brook trout and groundwater. Groundwater contributions to stream flow are dependent on the
surficial geology of the catchment. Catchments possessing a thick overburden have a large water storage capacity
and are capable of providing higher sustained base-flows during periods of low precipitation (Buttle et al. 2004, Buttle
and Eimers 2009). The ability of water to infiltrate and flow through the overburden is controlled by the type of surficial geology with course-grained substrate having a greater infiltration rate than fine-grained substrate (Neff et al. 2005). The type of surficial geology and its contribution to base flow is incorporated into the BFI score for each catchment. Therefore, the overburden thickness and the BFI score incorporate different, but related, aspects of the geology-groundwater connection. Previous studies have also found connections between surficial geology and brook trout occurrence (Witzel and MacCrimmon 1983; Stanfield et al. 2006; McKenna and Johnson 2011; DeWeber and Wagner 2015). For example, DeWeber and Wagner (2015) found that brook trout were positively related to areas of high soil permeability (i.e., course-textured sediment). In addition to groundwater, the surficial geology also influences sediment transport into rivers with areas of fine-textured sediment (low BFI) providing poor brook trout habitat because fine sediment negatively impacts brook trout spawning and feeding success (Argent and Flebbe 1999; Sweka and Hartman 2001).

Brook trout preferred forested landscapes and the proportion of forest was the most consistently selected landscape variable among the distribution models. The positive relationship between brook trout and forested area is consistent with other brook trout distribution studies (Hudy et al. 2008; Wagner et al. 2013; Kanno et al. 2015b). Wagner et al. (2013) found that brook trout occurrence was highest when >60% of the upstream network catchment was forested. Forested riparian vegetation shades streams from solar radiation, which leads to a decrease in diurnal temperature fluctuations and lower summer maximum stream temperatures (Barton et al. 1985; Johnson and Jones 2000; Cross et al. 2013). The cooler water temperatures of shaded streams provides favorable thermal habitat for brook trout. In contrast, human development of the landscape generally leads to the degradation of stream habitat due to increased stream temperatures (Nelson and Palmer 2007), increased sedimentation (Soulsby et al. 2001; Sutherland et al. 2002), altered stream flows (Paul and Meyer 2001; O’Driscoll et al. 2010) and pollution (Paul and Meyer 2001; Hatt et al. 2004). Previous studies have found that brook trout are negatively associated to agriculture and urban development (Stanfield et al. 2006; Wenger et al. 2011; Wagner et al. 2013; DeWeber and Wagner 2015). Only the logistic regression model and support vector machine model selected agriculture and community infrastructure landscape variables. The omission of human development metrics from the majority of the distribution models does not imply human development is innocuous. Instead, the models imply the same general relationship whereby brook trout are more likely to be found in areas of forested vegetation and less likely to be found in areas of human development.

The probability of brook trout occurrence was negatively related to the maximum monthly air temperature. Air temperature is commonly used as a surrogate for stream temperature because there is often insufficient information on stream temperature (McKenna and Johnson 2011; Wenger et al. 2011) Air and stream temperature are positively correlated, which makes air temperature a general indicator of stream temperature (Stoneman and Jones 1996, Chu et al. 2009, 2010). Brook trout are a coldwater species that are intolerant to warm stream temperatures (Wehrly et al. 2007). Elevated stream temperatures have been found to affect brook trout in a variety of ways, such as delaying reproductive timing (Warren et al. 2012) and increasing mortality (Hakala and Hartman 2004; Kanno et al. 2015a). Future climate change will only exacerbate these negative impacts and further restrict the range of brook trout (Chu et al. 2005; Wenger et al. 2011; Isaak and Rieman 2013).

The current distribution of brook trout is much more restricted than the hindcasted distribution, which suggests that brook trout have lost habitat within the Lake Simcoe watershed due to landscape alterations (i.e., deforestation). The current distribution of brook trout shows a clear retreat into the headwater regions (e.g., the Pefferlaw River; Figure 25–30). The hindcast distribution models simulate what the brook trout distribution would have been when the entire Lake Simcoe watershed was forested prior to human development. We acknowledge that our assumption of a forested landscape is a simplistic representation of past conditions, but it provides us with a reasonable means to assess brook trout distribution changes between past and current conditions. Hindcast modelling has also been used to predict the effectiveness of stream conservation and restoration efforts (Maire et al. 2015). For brook trout, a common approach to stream restoration is to increase the riparian vegetation through reforestation efforts. The results of our hindcast modelling can be used to help prioritize areas for riparian reforestation and predict how these efforts may impact brook trout populations. Stream reaches that responded positively to the increase in forested area are likely good candidates for riparian reforestation efforts. In addition, scenario models could be developed to understand how other changes to the landscape could influence brook trout habitat (e.g., urban sprawl).
Management implications

Brook trout are considered by the Lake Simcoe Protection Plan as indicators of water quality in tributary habitats because of their sensitivity to environmental conditions, such as temperature and dissolved oxygen (Ontario Ministry of the Environment 2009). The Lake Simcoe Fish Community Objectives recognizes the vulnerability of brook trout and specifies that existing brook trout populations be protected and coldwater stream habitat be restored where possible (OMNR 2011). The majority of brook trout in the Lake Simcoe watershed reside in headwater regions and conservation priority should be directed towards those headwater regions with brook trout present (e.g., the Pefferlaw River). Headwater regions provide valuable habitat for brook trout during all life history phases (spawning-adult; Witzel and MacCrimmon 1983; Petty et al. 2005; Kanno et al. 2015b). Most of the data on brook trout in the Lake Simcoe watershed has come from adult fish. However, there were some records for young-of-the-year (YOY) brook trout and they are also found in headwater regions (Figure 31). The overlap between adults and YOY qualitatively supports the use of headwater streams by various brook trout life-stages in the Lake Simcoe watershed.

The importance of headwater streams to the maintenance of brook trout populations in the Lake Simcoe watershed becomes even more apparent when considering the future temperature increases associated with climate change. The stream temperatures of Lake Simcoe tributaries are predicted to increase by as much as 1.3°C by 2100 (Chu 2011). This increase in stream temperatures will lead to significant range retractions for brook trout and other coldwater fish (Chu et al. 2005; Chu 2011). The range of brook trout will become increasingly restricted to the headwater regions as marginal habitat downstream becomes too warm (Meisner 1990a, 1990b; Wenger et al. 2011). Therefore, headwater habitats will be crucial to the current and future maintenance of brook trout populations in the Lake Simcoe watershed.

The restoration of stream habitat is also crucial to the maintenance of existing brook trout populations. Brook trout have lost a significant amount of habitat in the Lake Simcoe watershed (see hindcast predictions: Figure 25–30) and the future restriction of brook trout into headwater streams will only further fragment brook trout populations making them more susceptible to extirpation (Roberts et al. 2013). Most of the habitat variables included in the distribution models were geomorphological and cannot be improved through restoration or management actions (e.g., geology). However, the base-flow index and treed area do provide some possible targets for habitat restoration. Brook trout prefer streams with high base-flows and restoration efforts can focus on maintaining/improving base-flows by reducing water withdrawals and preventing the development of recharge areas. The Lake Simcoe Regional Conservation Authority has taken the first step in protecting recharge areas through the creation of a guidance document for development in recharge areas (LSRCA 2014). In addition, efforts to improve riparian vegetation through reforestation efforts will help remediate impaired brook trout habitat by decreasing stream temperatures and stabilizing banks. Together, these restoration efforts can help to increase brook trout habitat, reduce fragmentation and buffer populations from the effects of climate change.

When planning conservation initiatives, manager’s need to also consider the hierarchical nature of river systems whereby ecological processes operate at different spatial and temporal scales (Frissell et al. 1986). The habitat variables retained during model selection included a variety of spatial scales from the catchment down to the reach. Conservation projects are generally funded at a reach scale, but most ecological processes of interest operate at a larger spatial scale (e.g., sedimentation; Allan et al. 1997). Therefore, an integrated catchment scale approach to stream conservation should be used when planning conservation initiatives to ensure that reach scale efforts make a meaningful contribution to the functionality of the overall catchment (Allan et al. 1997; Dudgeon et al. 2006).

Advancing the distribution models

The distribution models performed well in the Lake Simcoe watershed and there was no obvious best performing model. However, there is still uncertainty about the predictions and disagreement among models (see Figure 17). A field validation of the model predictions is required to rigorously assess the predictions and to further evaluate model performance. Few studies have incorporated field validations as part of the model evaluation process (but see McKenna and Johnson 2011). Distribution models are meant to provide contemporary distribution estimates and field validations test model predictions on contemporary collections. Therefore, a prudent approach would be to use a combination of existing and field collected data to validate model predictions.
Figure 31. Location of brook trout young of the year in the Lake Simcoe watershed (red points).
The distribution models predict the presence or absence of brook trout around Lake Simcoe, which is a simplistic representation of brook trout occupancy. Brook trout habitat use can change as they develop from hatching to sexual maturity (i.e., ontogenic niche shift). Brook trout spawn in the gravel substrate of headwater streams and the young often remain in the general proximity of the spawning grounds (Petty et al. 2005). The sexually mature adults display greater dispersal ability and will often change position within a stream system depending on the season (Gowan and Fausch 1996; Curry et al. 2002; Petty et al. 2005). The brook trout distribution models for Lake Simcoe are based primarily on records of adult fish and an assumption of these models is that the distribution of adults will sufficiently describe the total distribution of brook trout during all life-stages. This assumption is implicit in the vast majority of brook trout distribution models because there is often a lack of data on the life-stage of captured individuals, including for the Lake Simcoe brook trout data (Stanfield et al. 2006; McKenna and Johnson 2011; Wenger et al. 2011; Wagner et al. 2013; DeWeber and Wagner 2015; Kanno et al. 2015b). The consistent identification of brook trout young-of-the-year during field sampling would allow separate distribution models to be built describing the spawning/juvenile habitat and the adult habitat. These life-stage specific distribution models would help fisheries managers target conservation efforts towards areas important for sensitive or limiting life-stages. For example, juvenile recruitment could be low in a system due to poor spawning success and a manager may want to conserve/improve spawning habitat. In order to do so, a manager would need to know the spawning sites of brook trout in the system.

The distribution models reflect the summer distribution of brook trout in the Lake Simcoe watershed. The brook trout data were collected during the summer months when the distribution of brook trout is limited to areas of shade, deep pools, or groundwater discharge due to high summer stream temperatures (Johnson and Dropkin 1996; Petty et al. 2005). Furthermore, the maximum monthly air temperature was used as a habitat variable in the distribution models. As a result, the distribution models do not necessarily reflect the full distribution of brook trout in the watershed. Brook trout are known make seasonal movements up and down stream systems depending on habitat and climatic conditions and will often occupy downstream stretches of river during the winter months (Curry et al. 2002). Distribution models rarely attempt to incorporate seasonal movements. However, these seasonal movements are essential to understanding the ecology and habitat use of brook trout in the Lake Simcoe watershed. The downstream movement of brook trout in the winter would result in a larger winter distribution and highlight barriers to movement. The collection of winter electrofishing data would provide the needed data to predict the winter distribution of brook trout.

Dams and other barriers can fragment a stream system by impeding or blocking brook trout movement. The habitat fragmentation caused by barriers can impact brook trout populations in a variety of ways, such as by reducing gene flow (Whiteley et al. 2013), increasing the probability of extirpation (Morita and Yamamoto 2002; Letcher et al. 2007) and exclusion from suitable habitat (Poplar-Jeffers et al. 2009). The distribution models for the Lake Simcoe watershed did not include the presence of stream barriers and, as a result, areas predicted to have suitable habitat may not possess brook trout. Few studies have incorporated stream barriers into their distribution models because there is often a lack of knowledge on the location of barriers and the amount of connectivity they provide (see Mahlum et al. 2014). Knowledge of barrier position in the Lake Simcoe watershed can help managers identify areas with suitable brook trout habitat that may benefit from barrier remediation efforts. Brook trout have strong dispersal capabilities and have been found to quickly recolonize areas after a disturbance (Roghair et al. 2002; Roghair and Dolfoff 2005). These traits make brook trout responsive to barrier removal and remediation efforts.

The distribution of brook trout can be shaped by interactions with non-native salmon (Stanfield et al. 2006; Wagner et al. 2013). Non-native salmon, such as rainbow trout (Oncorhynchus mykiss) and brown trout (Salmo trutta), have been introduced to stream systems throughout Ontario and are now ubiquitous within the Great Lakes Basin, less so inland (Crawford 2001). These non-native salmon require habitat similar to brook trout, which leads to niche overlap and competition. Non-native salmon tend to dominate during interactions with brook trout, particularly at higher water temperatures, leading to the competitive exclusion of brook trout from suitable habitat (Fausch and White 1981; Larson and Moore 1985; Ohlund et al. 2008; Thibault and Dodson 2013). This competitive exclusion has been shown to negatively influence the distribution of brook trout by restricting brook trout to headwater areas (Wagner et al. 2013). Stanfield et al. (2006) found that brook trout in northern Lake Ontario tributaries preferred stream reaches located in headwater areas that were absent of non-native salmon. The presence of non-native salmon in the Lake Simcoe watershed is limited relative to other areas in Ontario. Brown trout were located in Hawkestone Creek (1 site), the Pefferlaw River (2 sites) and Uxbridge Brook (10 sites). Rainbow trout were located in Bluff's Creek (3 sites), the
Pefferlaw River (1 site) and Uxbridge Brook (4 sites). Because of their limited distribution, we do not expect non-native salmon to greatly impact our predicted distribution of brook trout. However, this may not be the case if these distribution models are used for other areas in Ontario.

Electrofishing has an imperfect detection probability and the lack of brook trout in a sample does not guarantee that brook trout are absent from a site. An imperfect detection probability can lead to false negatives and the underestimation of a species distribution (DeWeber and Wagner 2015). The detection probability can even vary among sites due to differences in the stream habitat being sampled (Temple and Pearsons 2007; Hense et al. 2010; Wagner et al. 2014). For example, Hense et al. (2010) found that the detection probability of brook trout was related to stream width, gradient and conductivity. We did not possess the appropriate data to properly control for the imperfect detection of brook trout in our distribution models. However, we do not expect this to have a large impact on our predicted distributions because brook trout generally have a high detection probability when using electrofishing methods (0.72–0.99; Hense et al. 2010; Wagner et al. 2013, 2014; Kanno et al. 2015b). The site specific detection probability of brook trout in the Lake Simcoe watershed could be incorporated into future distribution models if three-pass electrofishing was conducted at a few sites or if all sampling sites were visited on multiple occasions (the former is preferable).

Modelling species abundance can provide valuable information not provided by presence-absence models, such as the conservation status (Gregory et al. 2004; Johnston et al. 2013), extinction risk (O’Grady et al. 2004) and population dynamics of a species (Letcher et al. 2007; Kanno et al. 2015a). Furthermore, abundance data can also be used to improve the fit of species distribution models (Howard et al. 2014). We attempted to model brook trout abundance in the Lake Simcoe watershed using catch per unit effort (CPUE) data. However, the available data did not contain sufficient effort information and the final CPUE dataset contained too few samples to build a reliable predictive model (70 samples; Figure 32). For example, we tried general linear, general additive and boosted regression tree models and neither of these models provided an $R^2 > 0.3$. The performance of the abundance models can be improved in the future by increasing the sample size. This would require more reliable and consistent information on the effort used to catch individuals. A more interpretable and commonly used measure of abundance would be the density of fish, which requires data on the area sampled (Petty et al. 2005; Stanfield et al. 2006; Steen et al. 2008; McKenna and Johnson 2011). The HabProgs database often lacked information on the length and width of a sampled stream and this prevented any calculations of brook trout density.

Better information on groundwater quality and quantity would be very useful for developing models that are more accurate and have better spatial resolution. Brook trout distribution is intimately related to cold groundwater sources, both point (e.g., seeps) and diffusely over large areas. More sophisticated models could be developed if there was better information on geology characteristics.
Figure 32. Brook trout catch per unit effort (CPUE: # fish caught/minute) in the Lake Simcoe watershed.
Recommendations

- Conserve the cold headwater areas where brook trout are present and restore marginal habitat (reforestation and shade riparian) to increase available brook trout habitat and improve connectivity.

- More long-term monitoring sites are required to better assess changes in brook trout population dynamics (see Wagner et al. 2014). Monitoring sites should be located along a gradient of brook trout habitat suitability because this will help in the early identification of distribution changes.

- Distribution model agreement and probability of occurrence mapping can be used to help target and set priorities for targeted brook trout inventory.

- Improve electrofishing records (information management) to allow for the estimation of brook trout abundance at all sites by recording accurate shocker seconds and consistently collecting information on the stream area sampled.

- The sampling methodology used to sample each site should be recorded to ensure all data is collected using a standard approach and to indicate when a site was sampled using an alternative methodology.

- Record the life-stage of the fish captured when taking individual and bulk measurements. This will allow the data to be easily separated by life-stage.

- Sample streams during the winter season to provide information on the winter distribution of brook trout in the system.

- All dams and barriers need to be mapped in the watershed and the degree of connectivity each barrier provides assessed.

- The distribution models could be coupled with groundwater discharge and stream temperature models to better predict areas of occupation.

- The predictions of the distribution models need to be field validated to properly assess their effectiveness.
References


Lake Simcoe Regional Conservation Authority. 2014. Guidance for the protection and restoration of significant groundwater recharge areas (SGRAs) in the Lake Simcoe watershed. Pages 1–16.


Ohlund, G., F. Nordwall, E. Degerman and T. Eriksson. 2008. Life history and large-scale habitat use of brown trout (Salmo trutta) and brook trout (Salvelinus fontinalis) - Implications for species replacement patterns. Canadian Journal of Fisheries and Aquatic Sciences 65:633–644.


Thuiller, W., B. Lafourcade and M. Araújo. 2010. The presentation manual for BIOMOD. Laboratoire d’Ecologie Alpine, Université Joseph Fourier, Grenoble, France.


Software


