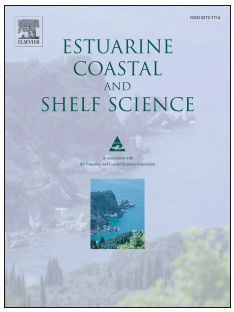


Deep learning habitat modeling for moving organisms in rapidly changing estuarine environments: A case of two fishes

Guillaume GuÉnard, Jean Morin, Pascal Matte, Yves Secretan, Eliane Valiquette, Marc Mingelbier

PII: S0272-7714(19)30427-5

DOI: https://doi.org/10.1016/j.ecss.2020.106713

Reference: YECSS 106713

To appear in: *Estuarine, Coastal and Shelf Science*

Received Date: 1 May 2019

Revised Date: 12 March 2020

Accepted Date: 16 March 2020

Please cite this article as: GuÉnard, G., Morin, J., Matte, P., Secretan, Y., Valiquette, E., Mingelbier, M., Deep learning habitat modeling for moving organisms in rapidly changing estuarine environments: A case of two fishes, *Estuarine, Coastal and Shelf Science* (2020), doi: https://doi.org/10.1016/j.ecss.2020.106713.

Deep learning habitat modeling for moving organisms in rapidly changing estuarine environments: a case of two fishes

Guillaume Guénard*,[1], Jean Morin[1], Pascal Matte[2], Yves Secretan[3], Eliane Valiquette[4], and Marc Mingelbier[4]

March 18, 2020

1. National Hydrological Service, Environment and Climate Change Canada, 1550 avenue d'Estimauville, Quebec City QC, G1J 5E9 Canada

2. Meteorological Research Division, Environment and Climate Change Canada, 1550 avenue d'Estimauville, Quebec City QC, G1J 5E9 Canada

3. Centre Eau Terre Environnement, Institut National de la Recherche Scientifique, 490 rue de la Couronne, Quebec City QC, G1K 9A9 Canada

4. Direction de l'expertise sur la faune aquatique, Ministère des Forêts, de la Faune et des Parcs (Québec), 880 Ch. Ste-Foy, Quebec City QC, G1S 2L2 Canada

Running head: Deep learning estuarine habitat modeling

*corresponding author, email: guillaume.guenard@umontreal.ca, tel.: +1-514-707-9569

Keywords: Numerical habitat model, Artificial neural network, Tide, Ecohydraulics, Two-dimensional hydraulic model, Numerical terrain model

Abstract

Modeling the spatial distribution of mobile organisms under rapidly changing environmental conditions is a challenging endeavor that has to be undertaken whenever the impacts of alterations have to be assessed in dynamic scenarios. We modeled habitat suitability for Lake sturgeon (*Acipenser fulvescens*) and White perch (*Morone americana*, both had have been followed by hydro-acoustic telemetry) in an estuarine river section with rapidly changing tidal and hydrodynamic conditions using deep feed-forward Artificial Neural Networks ANN). Descriptors used were of many types: intrinsic features (species, sexual maturity and gender, and individual character), terrain features, hydraulic and tidal conditions, and time. A set of ANN models with varying degree of complexity, in terms of their number of hidden layers, number of nodes per layers, and regularization parameters, were tried and evaluated using cross-validation. The best model has three layers with 100, 50, and 20 nodes and classified 94.0% of observations as presence (and 60.6% of pseudo absences as absences, overall correct classification: 77.3%) during the trials. The study highlights that tidal and hydraulic models, coupled with acoustic telemetry and machine learning, can be used to predict the spatial distribution of mobile organisms even in extremely variable ecosystems such as estuaries.

Introduction

As with most rapidly changing system (e.g., hydropeaking, large alluvial floodplain, intermittent rivers), estuaries are significant challenges to habitat modeling. The tidal conditions prevailing in temperate estuaries involve marked fluctuations of water level and current velocity, accompanied by changes in wet/dry areas and current reversals, substantially increasing the complexity of obtaining dependable

estimates of the physical conditions (e.g. Skov et al., 2008; Spruzen et al., 2008; Sagarese et al., 2014). In the present paper, the word "estuaries" will refer specifically to temperate estuaries or to tropical estuaries having no barrier reef, and which experience tidal conditions that are typical of temperate ones. These fluctuating conditions force mobile animals to travel more often about habitats, which are successively made suitable and unsuitable to them to a greater extent than they typically would in a non-tidal system (estuarine species are especially well-adapted to these conditions, see Gibson, 1993). Estuaries often harbor rich and diversified ecosystems that are oftentimes harshly impacted by human activities (Roessig et al., 2004; Lotze et al., 2006). There is thus a clear incentive in unfolding the best of today's scientific and technical knowledge to take up the challenge of numerical habitat modeling in estuarine conditions.

Two-dimensional (2D) hydraulic models (2DHM) have proven their worth for habitat modeling in rivers (e.g., Guay et al., 2000; Mingelbier et al., 2008; Morin et al., 2016; Capra et al., 2017; Foubert et al., 2019). In a nutshell, these models estimate 2D (scalar) fields of water level, vector fields of current velocity, and other related physical quantities (e.g., water depth, shear velocity, Froude number). To achieve that goal, 2DHM require maps of riverbed elevation and roughness, and user-specified boundary conditions (i.e., flow and/or water level), while making a set of assumptions about fluid mechanics (Heniche et al., 2006). 2DHM most commonly implement the 2D depth-integrated Navier-Stokes equations on a 2D grid domain. The estimated water level, current velocity, direction, and other physical variables output by the 2DHM, combined with other terrain features (e.g., substrate composition, bottom slope), can be used as descriptors of the river habitat. In addition to 2DHM, three-dimensional hydraulic models also exist (e.g., the MARS3D model: Dumas and Langlois, 2009; `https://wwz.ifremer.fr/mars3d/Presentation`) but have not yet been implemented in the St. Lawrence fluvial estuary (SLFE).

In rivers and estuaries, tides are externally forced by multi-scale and non-stationary signals interacting in a nonlinear manner with bottom friction and basin

topography. They are influenced, on the one hand, by river flow that varies over meteorological, seasonal, and longer-term time scales. On the other hand, they are driven by the astronomical tides propagating from the sea along with other meteorological and oceanographic signals, such as storm surges and sea-level rise. The analysis of non-stationary tides and currents requires well-adapted analytical tools relying on detailed *in situ* data. For example, Matte et al. (2014c) have developed an improved harmonic regression method predicting tidal waves as they interact with river discharge, and applied it to the SLFE. Furthermore, studies conducted by Matte et al. (2017a) and Matte et al. (2017b) in the SLFE, supported by extensive field campaigns (Matte et al., 2014b,a), showed that information about the tide, *in situ* water currents, and depths is instrumental in implementing hydraulic models in estuaries.

A numerical habitat model (NHM) uses descriptors of environmental conditions to assess whether and to what extent an area is suitable for a species (Guay et al., 2000; Boisclair, 2001). Because the life history of an organism is composed of many steps (e.g., larval, juvenile, reproduction, wintering), involving their particular needs, NHM are, in common practice, targeted at a particular age class and time of the year. NHM are often implemented as regression models representing mean density or probability of occurrence. Recent advances in numerical modeling methods (e.g., machine learning) now provide us with efficient new means for representing non-linear relationships and intricate interactions among variables (Lek et al., 1996; Brosse et al., 1999; Olden and Jackson, 2001; Quetglas et al., 2011). Among these modeling methods, artificial neural network (ANN) allows one to model such complexity. More importantly, ANN achieves that goal without the requirement of explicitly defining a suite of non-linear (e.g., polynomials) and interaction terms. Better still, that adaptability of ANN enables a single NHM to implicitly represent the different life history steps of an organism using information on the latter's traits and status (e.g., age, size, sex), and time (e.g., season, time of day) and this without having to make assumptions about its specific behavior and traits. In fact, ANN have enough flexibility to incorporate multiple species, together

with their various life history steps, in a single NHM, provided they are given suitable descriptors; (e.g. Guénard et al., 2017) and sufficient data be available. Hence, functional traits may be shared among species and relationships between these traits and the responses of organisms to the environment may be partially redundant. Having a single model NHM spanning many species and their different life stages thus offers the advantage of maximizing information use while requiring less development efforts in comparison with that of developing suites of individual models.

The objective of the present study is to develop, test and show benefits of ANN using dynamic variables from 2DHM to model probability of presence for fish in a highly varying environmental setting associated with non-stationary estuarine tide. The first steps of this modeling exercise have already been achieved by Matte et al. (2014b; 2014a; 2014c; 2017a; 2017b), who modeled tidal waves that are used as boundary conditions for the modeled area (see subsection *Time steps and boundary conditions* below for details), performed real-time field measurement of water levels and current velocities, and applied 2DHM in and around the study area. The main goal of the present study will focus on using the information about the physical conditions collected in these prior studies to implement a single temporally-explicit, spatially-explicit, and individually-based probabilistic NHM. This model, which is based on multiple-layer (deep) ANN, will handle any variation potentially associated with the different life-history steps of the two fish species in an implicit manner.

Material and Methods

Case study areas

Hydraulic simulations

The hydraulic simulation area was a 140 km long section of the St-Lawrence river starting from an imaginary line drawn from Neuville (Quebec, Canada; WGS84: +46.6960, −71.5727) on the north-western shore to Saint-Antoine-de-Tilly (Quebec,

Canada; WGS84: +46.6681, −71.5546) on the south-eastern shore and ending in the SLFE on a line stretching from Les Éboulements (Quebec, Canada; WGS84: +47.48, −70.2579) on the north-western coast to Rivière-Ouelle (Quebec, Canada; WGS84: +47.4625, −70.0278) on the south-eastern coast (Fig. 1). The information on terrain elevation needed to implement a digital elevation model (DEM) of the area was available from earlier bathymetric and topographic surveys (Matte et al., 2017a,b). Similarly, substrate data were used to define regions of homogeneous substrate composition, converted into Manning's coefficients using the grain size classification and formulation developed by Morin et al. (2000). The surface of the modeled area was approximately 1 800 km^2.

Fish tracking survey

The fish tracking area is a river section within the hydraulic simulation area, which is confined approximately 2 km upstream and 2 km downstream of the road bridge connecting Île d'Orléans to the continent (French: *pont de l'Île d'Orléans*, hereafter abbreviated PIO; Fig. 1). It is strictly freshwater, located downstream of Quebec City's harbor, between Île d'Orléans and Beauport borough (Quebec city) on the north shore, and is among the most energetic regions of the St. Lawrence, with tidal a range exceeding 6 m under large spring tide conditions, leading to peaks in tidal discharges of up to five times the daily average discharge of the river ($\approx 12\,200\,m^3\,s^{-1}$ at Quebec City) in both the upstream and downstream directions (Matte et al., 2017a).

The field telemetry surveys were performed in years 2012 and 2013 for developing tools to assess changes in habitat quality related to the anticipated changes in physical conditions associated with planed road bridge construction works. The complete procedure is detailed in Valiquette et al. (2016). Briefly, 33 Lake sturgeon (*Acipenser fulvescens*) and 15 White perch (*Morone americana*) were implanted with Vemco® hydro-acoustic transmitters model V16 and V9, respectively. Fish length, weight, and sex were taken during captivity. Tracking of the marked individuals was achieved in 2013 using a Vemco Positioning System (VPS) array of

Table 1: Species features

Characteristic (typical)		Lake sturgeon	White perch
Adult total length		92 – 123 cm	20 – 25 cm
Adult weight		4.5 – 36 kg	250 – 500 g
Age at sexual maturity		12 – 23 y	3 – 7 y
Egg per female		100 000 – 900 000	20 000 – 300 000
Spawning	Time	early May – late June	early May – late June
	Temperature	13 – 18°C	10 – 20°C
	Depth	0.6 – 4.6 m	0 – 3.7 m
Cause of threat		Overfishing, pollution	Spawning ground perturbation

24 Vemco® VR2W hydro-acoustic receivers providing fish positioning. We deemed observations with Hyperbolic Positioning Error (HPE) > 100 m as well as that under a shallow water threshold of 30 cm too inaccurate and discarded them from further analyses. Lake sturgeon is listed as a vulnerable species in Canada (COSEWIC, 2017). White perch is an abundant fish in the study area (Valiquette et al., 2016; Table 1).

The area used for the fish tracking survey was considerably smaller than that for hydraulic simulations (5.36 km^2 vs. 1 800 km^2; see Fig. 1). The following reasons underpin that choice of a larger area. Firstly, only a small portion ($\leq 10\%$) of the total St. Lawrence River flow is channeled through the study area (in "Chenal de l'Île d'Orléans"; CIO); the remaining flows in the channel south of Île d'Orléans (in "Chenal des Grands Bateaux"; CGB). Flow and tidal conditions affecting hydraulics in CGB have substantial effects on CIO; the former thus needed to be part of hydraulic simulations. Secondly, uncertainty may appear in the boundary conditions, whose values are assumed to be invariant along boundaries. Estimated hydraulics near the model boundaries are thus less accurate than farther away, inaccuracies averaging out as we move away from them. Given the complexity of the topography in the vicinity of Île d'Orléans and of the tidal conditions prevailing in the SLFE, the area for hydraulic simulations had to be much larger than that where the NHM was to be developed.

Modeling physical conditions

Hydraulic modeling

The hydraulic simulation area was discretized into a 2D finite-element mesh made of 106 520 finite elements involving a total of 216 913 nodes. The finite element used is called P1–P1isoP2 consists of a triangle made of six nodes: three nodes at the triangle's vertices and three nodes located at mid-distance along each of its three edges (see Heniche et al., 2006, for details). The elemental surface areas ranged from 595 cm^2 to 103 994 m^2, with a median size of 985 m^2. Elements size is chosen to match local complexity in both topography and hydraulic conditions, with increasingly small elements used when conditions are known or expected to be locally variable in space and a few larger elements for areas with uniform properties.

Time steps and boundary conditions

The dynamic nature of the physical conditions prevailing in the study area required performing non-stationary hydraulic simulations using short (3 min) time steps, in order to capture the system variability and ensure fast convergence. For the period of interest (i.e., 2013-04-30 23:36:00 through 2013-11-21 05:09:00), this represents a total of 98 051 time steps. Combined with the large size of the modeled area and the high mesh resolution, the computation time required for such a large number of time steps would have been impractical.

To circumvent that computational issue, we defined a set of synthetic hydraulic conditions meant to be representative of the river flows and estuarine tides encountered in the system. These synthetic conditions were substituted to those observed. To achieve such a synthesis, we first built two eight-day tidal height time series with 3 min time steps. Each series was meant to encompass a large range of tidal conditions, from neap to spring tides, observed in the SLFE for one of two scenarios of river flow. These two scenarios corresponded to river flows of 10 477 m^3s^{-1} (mean annual discharge) and 17 609 m^3s^{-1}(mean high discharge), with a total of 3 840 consecutive time steps per scenario, thus considerably reducing the

computational burden. With these scenarios, a sufficient range of tidal and flow conditions encountered during the period of study were covered. Hydraulic modeling was performed separately on these two synthetic time series. Hydraulic conditions at any given time during the study period were taken as those estimated for the synthetic condition that most closely matched that in the field. Matching was done with respect to similarity between measured and synthetic tidal height time series for a reference location, in terms of mean water level and tidal range. Once the best synthetic tidal cycle and time step was identified, the corresponding simulated 2D fields were attributed to that given time. This process was repeated separately for each tidal cycle within the period of study, allowing us to reconstruct the 2D hydraulic conditions for the entire period.

Tidal modeling

Values of tidal heights prescribed at the upstream and downstream boundaries of the 2DHM were obtained from the NS_TIDE model and software developed by Matte et al. (2013; see *supplementary material* for a description). We used the NS_TIDE spatial model of the SLFE to generate the aforementioned synthetic tidal height time series. Conditions of tidal range were specified at the same oceanic reference station (Sept-Îles, Quebec, Canada) following Matte et al., 2014c. We took tidal ranges increasing steadily from 1.3 m (first percentile) to 3.4 m (99^{th} percentile) over the eight-day synthetic period, thereby simulating a transition from a neap tide to a spring tide. Tidal heights at the upstream and downstream limits of the 2DHM area were reconstructed for each time step of the synthetic time series and for each of the two scenarios of river flow. Tidal heights were also calculated for the Vieux-Québec tide gauge (Department of Fisheries and Oceans Canada, Gauge number 3248; WGS84: +46.811 111, −71.201 944), the closest permanent gauge located within the hydraulic model area and a few kilometers upstream of our study area. This station is the local reference location that we used in selecting synthetic simulation results that most closely matched hydraulic conditions in the field.

Regular grid

While an unstructured finite-element mesh was necessary for efficient computation of the 2DHM, its elements have heterogeneous sizes that make them less suitable for NHM, for which a regular grid is preferred in practice (e.g. Morin et al., 2016). Therefore, we discretized the study area once more, but this time into a regular square grid composed of a total of 214 339 tiles with 25 m^2 surface area, in order to facilitate subsequent computations. We linearly interpolated information on hydraulics, which is available at the vertices of each triangular element of the mesh, onto the regular square grid from a plane defined by the three nearest nodes of the finite-element mesh. Fish tracking events, which are continuous in space and time, were pinned down to the regular grid (in space) and sampling times of the tidal time series from the Vieux-Québec tide gauge.

Fish tracking

In the present study we used the information associated with fish location at a particular time. The other type of information entailed by tracking –the information related to the direction and distance traveled by the fish during a given time period– will not be used here. While fish positioning brings information about the conditions sought after by the fish, it does not explicitly provide us with information about the conditions avoided by them. Nevertheless, the latter is required by common NHM approaches yielding probabilities of presence on site and the requiring information on conditions for both presence and absence.

Pseudo absence

A workaround in such circumstances is to supply the lack of absence data by taking a set of locations, called pseudo absences (PA), that are to be considered as absence data (VanDerWal et al., 2009; Barbet-Massin et al., 2012). To avoid confounding unsuitable habitats with that which cannot be sampled, PA must only be placed on locations where fish could have been tracked at sampling time. For our case study,

it means inside the area where transmitters could effectively be tracked by the VPS array at a given time. We thus chose PA in conjunction with individual fish detection events and selected them from pools of locations featuring conditions that were dissimilar to that of observation sites. Dissimilarity was measured using the Mahalanobis distance (Mahalanobis, 1936), calculated on the basis of the descriptors defined below. For any given fish tracking event, only grid points with a depth $\geqslant 30\,\text{cm}$ (the swallow water threshold) and a Mahalanobis distance > 1.96 from the conditions observed on the observation site (i.e., having a probability < 0.05 of belonging to the presence group under the multi-normality assumption) were in the PA pool, from which a single PA was drawn. For model testing purposes, we set aside six fish (three Lake sturgeons and three White perches) as the model's testing set and used the remaining data (30 Lake sturgeons and 12 White perches) as the model's training set.

Descriptors

There are several types of descriptors that can be derived from the data available for the present study, namely individual, temporal, tidal, terrain, and hydraulic descriptors. A total of 64 descriptors was defined; they are detailed below.

Individual descriptors

In most tracking data, many individuals are being followed and some individuals are being observed more often than others. Any individual organism has its own character, which may deviate from the mean behavior of the population. As a consequence, the most frequently observed individuals would make their particular behavioral traits appear more widespread than their actual importance in the population. To prevent such bias, we supplied the NHM with information on fish identity, which consisted in representing it as a fixed effect using contrast variables. Each contrast took the value $+1$ for the N_i data related to individual i and the value $-N_i/(N - N_i)$, where N is the total number of observations, for data related

to the other individuals in the sample. Contrasts defined in that manner are, by definition, centered on the value 0; providing the NHM with 0 for all contrasts thus allowed it to perform out-of-the-sample predictions.

Besides individual contrasts, descriptors were also used for representing the species (one binary descriptor taking the value 1 for Lake sturgeon and 0 for White perch) and reproductive status of the fish. The latter consisted of a pair of binary variables with $\{1, 0\}$ representing females, $\{0, 1\}$ males, and $\{0, 0\}$ for sexually indeterminate individuals.

Terrain and hydraulic descriptors

Terrain descriptors were the bottom slope, bottom curvature, substrate composition (i.e., percentages of each substrate class), and vegetation coverage (Morin et al., 2003), whereas hydraulic descriptors were the water depth, the mean current velocity (norm of the depth-averaged velocity vector), and the bottom slope in the flow direction (BSFD; for details on terrain and hydraulic descriptors calculation, see *Supplementary material*).

Tidal descriptors

We considered the possibility that fish have an anticipatory sense of the changing tide. Such a skill, whether it stems from a sense of timing, sensory cues, or other means, is potentially adaptive as it may enable fish to modulate their behavior and help them thrive. It is therefore expected to have evolved in estuarine species (Gibson, 1993). The function of the tidal height with time has a periodic character: individual tidal height values are thus not informative of local rates and directions (rising or falling) of change in the time series. To obtain a thorough representation of the status of the tide at any given moment, we paired the tidal height time series with a second time series of its first derivative with respect to time: the tidal rate (the instantaneous rate of change on tidal height, in $\mathrm{m\,h^{-1}}$, see *Supplementary material* for properties and computation details; Fig. 2).

Temporal descriptors

As mentioned earlier, fish positioning was performed automatically using acoustic telemetry, with observations occurring throughout different seasons and at different time of the day. We expected fish to behave differently at different times and thus provided the NHM with descriptors about the course of time, which we also expect to relate with other unmeasured variables such as daylight intensity and duration, and water temperature. We thus needed variables representing the passage of seasons and days unequivocally, with pairs of values that are never identical for different times of the same year. To achieve that goal, we used two pairs of quadrature descriptors: SeasonA and SeasonB to represent the passage of seasons (a proxy to light period, water temperature, and so on; Fig. 3), and CircadianA and CircadianB to represent that of daytime (Fig. 4; see *supplementary material* for properties and computation details).

Modeling

Modeling the probability of presence

As it is the case for most machine learning approaches, ANN is suitable in situations where copious amounts of data are available and the effects of the descriptors on the response are expected to be complex and tedious to define analytically. This is what we expect in the present case as habitat selection hinges on a suite of behavioral norms that are expected to vary among individuals of different species, sex, and reproductive status; and at different times, whereas acoustic telemetry allowed generous amounts of positioning information to be collected. We used a deep ANN to represent fish probability of presence on the grid points from the descriptors. Since ANN is becoming increasingly widespread and its thorough description would unacceptably lengthen the present paper, one is provided as *Supplementary material*.

Estimation and comparison

In machine learning, parameters such as those defining ANN structure and how weights are estimated are called "hyper parameters" (HP). Hence, the number of hidden layers and their numbers of node are HP, as well as the weight regularization norms and dropout rates (see *supplementary material* for details). We performed a grid search to estimate the best HP set. It consists in defining discrete sets of values deemed reasonable and perform cross-validation trials over these sets to find that yielding the NHM with the best predictive performance. Cross-validation folds were the individual fish: any given individual's probability of presence was thus predicted from data on the remaining samples. We performed the search over eight hidden layer configurations: $\{10\}$, $\{20\}$, $\{30\}$, $\{10, 5\}$, $\{25, 10\}$, $\{25, 10, 5\}$, $\{50, 25, 10\}$, or $\{100, 50, 25\}$; where $\{n_1, n_2, n_{...}, n_m\}$ represents a configuration with m hidden layers whose first layer has n_1 node, second layer n_2 nodes, etc., representing networks with 650, 1300, 1950, 705, 1885, 1940, 4785, or 12825 weights, respectively, five values for the L_1 norm: 1e-07, 1e-06, 1e-05, 1e-04, or 1e-03, and the same five values for the L_2 norm (total: 200 HP sets). For the sake of simplicity, we used constant dropout ratios of 20% for the input layer and 50% for the hidden layers, a single L_1, L_2 pair was used over all layers, and used the rectifier activation function for all hidden layers (see *Supplementary material* for details). We finally used the sigmoid (inverse logit) function for the model output as it is the canonical link function (sensu Hastie and Pregibon, 1991) for Bernoulli-distributed, absence / presence, response data. The resulting NHM therefore outputs (fitted or predicted) probabilities of presence.

We compared the deep ANN model developed in the present study with a more classical binary classification approach: logistic regression. To build that model, and for the sake of easing model comparison, we used the same cross-validation approach as for the deep ANN model to estimate the (L_1 and L_2) regularization parameters. For both the ANN and the logistic regression models, the decision threshold to predict fish presence was that maximizing the F1 score (Lipton et al., 2014).

Variable profiling

ANN models allows any given environmental variable to influence its response in different ways under different sets of conditions (i.e., values of the other variables). It is straightforward to display the NHM's response to the fluctuation of a given variable, by sweeping it over its entire range while keeping all other variables constant as a reference condition. Then, by carefully picking suitable reference conditions, one can obtain a variety of profiles showing a broad range of different model outputs. The set of all possible reference conditions can be regarded as a continuous space under which any given variable can be profiled. Here we used kmeans clustering to select a representative subset from that space. The number of reference conditions in that subset was estimated using Calinski's criterion (see Legendre and Legendre, 2012, for a desccription). To calculate the kmeans clustering, we selected all the descriptors besides the dummy variables representing the species and genders, and the individual contrasts. Variable profile were calculated separately for each species and gender (female, male, and indeterminate), over the entire range of each descriptor, and for the whole subset of reference conditions.

Habitat mapping

A widespread goal of NHM is to provide decision-makers with habitat quality maps allowing them to plan construction work while abiding to laws and regulations about imperiled species and meet sustainable development objectives. ANN models can generate a staggering variety of predictions that is challenging for the human mind to integrate (see the 288 habitat quality maps in *Supplementary material* as an example). To help us synthesize that information, we proposed to use presence proportions maps. A presence proportion map gives, for every grid location, the proportion of the time a presence is predicted (i.e., exceeds the model's threshold for presence) during a given time period. Proportions vary between 0 (for systematic unsuitability) and 1 (for systematic suitability). We calculated the

proportions on the basis of the number of times predictions could be made. Indeed, proportion could not be calculated where depth was systematically outside the range of the model ($z < 30\,\text{cm}$ or $z > 30\,\text{m}$). Presence proportion maps were calculated for female and male of either species for two periods: during both species spawning period (May 01 to June 30) and after spawning (July 01 to Nov. 21).

Software

Hydraulic modeling was performed using H2D2 (Heniche et al., 2006; Matte et al., 2017a; `http://www.gre-ehn.ete.inrs.ca/H2D2`), a software solving the 2D Navier-Stokes equations over a finite-element discretized domain, and which include a drying-wetting model for the treatment of areas becoming wet or dry as water level rises and falls. All data manipulations were made using the R language and environment (R Core Team, 2017) and contributed package available from the Comprehensive R Archive Network (CRAN; `https://cran.r-project.org`). Packages DBI (R Special Interest Group on Databases (R-SIG-DB) et al., 2016), ROracle (Mukhin et al., 2016), and RSQLite (Wickham et al., 2014) were used for database transactions. Packages raster (Hijmans, 2016), rgdal (Bivand et al., 2017), and sp (Pebesma and Bivand, 2005) were used to manipulate geographic information, with the QGIS software (v2.18, `https://www.qgis.org`) used for geographic data visualization. Package h2o (The H2O.ai team, 2017) was used for communication with the h2o software (`https://www.h2o.ai`), whose module h2o.deeplearning (Candel et al., 2018) was used to estimate ANN.

Results

Telemetry study

In 2013, a total of 36158 detection events were recorded by the VPS array, with numbers of detection events per fish varying between 2 and 4861. The data set with generated PA set had a sample size of 72316. The three Lake sturgeons and three

Table 2: Ranges of the continuous descriptors that were experienced by the fish of either species (Presence) or drawn as pseudo absences (Absence), and available somewhere on the sampling grid (Available).

Descriptor	Units	Lake Sturgeon		White perch		Available
		Presence	Absence	Presence	Absence	
Slope		[0.00, 0.23]	[0.00, 0.46]	[0.00, 0.22]	[0.00, 0.46]	[0.00, 0.58]
Curvature		[0.00, 0.08]	[0.00, 0.07]	[0.00, 0.08]	[0.00, 0.07]	[0.00, 0.08]
Clay/silt	%	[00.0, 94.6]	[00.0, 94.6]	[00.0, 94.6]	[00.0, 94.6]	[00.0, 94.6]
Sand	%	[03.8, 70.0]	[03.8, 70.0]	[03.8, 70.0]	[03.8, 70.0]	[03.8, 70.0]
Gravel	%	[01.5, 60.0]	[01.5, 60.0]	[01.5, 60.0]	[01.5, 60.0]	[01.5, 60.0]
Pebble	%	[00.0, 43.2]	[00.0, 43.2]	[00.0, 43.2]	[00.0, 43.2]	[00.0, 43.2]
Cobble	%	[00.0, 06.8]	[00.0, 06.8]	[00.0, 06.8]	[00.0, 06.8]	[00.0, 06.8]
Boulder	%	[00.0, 05.8]	[00.0, 10.0]	[00.0, 00.0]	[00.0, 05.8]	[00.0, 10.0]
Velocity	$\mathrm{m\,s^{-1}}$	[0.0, 0.9]	[0.0, 1.0]	[0.0, 0.9]	[0.0, 1.0]	[0.0, 1.2]
Depth*	m	[0.8, 23.9]	[0.3, 24.4]	[0.3, 22.9]	[0.3, 23.9]	[0.3, 25.1]
Sheer Vel.	$\mathrm{m\,s^{-1}}$	[0.0, 0.0]	[0.0, 0.2]	[0.0, 0.1]	[0.0, 0.2]	[0.0, 0.3]
BSFD**		[−0.2, 0.2]	[−0.2, 0.2]	[−0.1, 0.1]	[−0.1, 0.1]	[−0.5, 0.3]
Tide height	m	[−2.0, 4.8]	[−2.0, 4.8]	[−1.9, 4.0]	[−1.9, 4.0]	[−2.0, 4.8]
Tide rate	$\mathrm{m\,h^{-1}}$	[−1.5, 2.6]	[−1.5, 2.6]	[−1.4, 2.5]	[−1.4, 2.5]	[−1.5, 2.6]

* Locations with depths < 30 cm were excluded (see text for rationales).
** Bottom slope in the flow direction

White perches set aside for model testing purposes made up a total of 3178 observations. The number of observations available for model training (estimate weights and HP) was thus 69138.

Fish were tracked during a period of 204 days spanning from 2013-04-30 23:36:00 to 2013-11-21 05:09:00. They were observed on sites covering a large portion of the range of environmental conditions found on the sampling grid and at all tidal heights and rates that were recorded in the study area (Table 2). The ranges for PA were typically wider than that of observations, albeit the rarest conditions were not necessarily represented as the PA were randomly drawn.

Fish distribution model

The best HP set found has three hidden layers having 100, 50, and 25 nodes each, whereas the L_1 and L_2 regularization parameters were 1e-03 and 1e-04, respectively. That HP combination was associated with a correct classification rate of 78.4%

(89.8% for presences and 67.0% for absences) during cross-validation (decision threshold: 0.4143; Fig. 5). The NHM estimated with the whole training data set had a correct classification rate of 77.3% (94.0% for presences and 60.6% for absences), whereas the same metrics calculated on the testing data set were 69.7% (86.7% for presences and 52.7% for absences). The median NHM response for sites where fish were observed was 0.70 (with the 5^{th} and 95^{th} percentiles being 0.26 and 0.86, respectively) whereas that for sites where fish were not observed was 0.17 (with the 5^{th} and 95^{th} percentiles being 0.00 and 0.71, respectively). By comparison, we found the best L_1 and L_2 for the same two-species logistic regression model to be both 1e-01, yielding a correct classification rate of 72.5% (90.1% for presences and 54.8% for absences) during cross-validation, 70.8% (84.6% for presences and 56.9% for absences) when applied to its training data, and 64.8% (74.0% for presences and 55.5% for absences) when applied to the testing data. The neural network-based NHM thus outperformed the logistic NHM.

Variable profiling

During kmeans clustering, the optimal number of reference condition clusters estimated from the Calinski criterion was 12. Among-cluster variation was mainly associated with substrate composition (see *Supplementary material*, Fig. 2 for details). The resulting 17 figures (*Supplementary material*, Figs. 3-19), each comprising six panes (two species times the three gender status) that themselves contained a curve for each cluster (a total of 1224 profiles were thus calculated), showing a broad array of different scenarios. Indeed, they were too numerous to be part of the present paper's main text (readers are referred to the *Supplementary material* for the details). For instances, a variable may appear to have an effect on the probability of presence only for a part of its range; this effect may only be apparent for specific reference conditions, with direction changing among the reference conditions. Also some variables may have similar profiles among species and genders (e.g., bottom slope; *Supplementary material*, Fig. 3) whereas some others may display more contrasting profiles among them (e.g., the percentage of

clay or silt in the sediment; *Supplementary material*, Fig. 6).

Habitat mapping

An apparent feature of the presence proportion maps is the general preference of both fishes for the area surrounding the actual bridge (Figs. 6 and 7). Lake sturgeon of either gender have very similar distributions, and this both during and after spawning. During spawning, both female and male Lake sturgeon appear to have a greater fondness toward the area located at the mouth of Montmorency river than after spawning. That area often is under waters that are too shallow for fish to be reliably detected by telemetry and predicted by the model, yet it appears to be preferred whenever such preference can be evidenced. That preference subsides after spawning, Lake sturgeon of either genders then appear to shift their distribution towards the middle of CIO, somewhat northerly to the actual road bridge.
In sharp opposition to Lake sturgeon, White perch do not appear to display any particular preference for the Montmorency river mouth area, be it during or after spawning. On the other hand, spawning female and, but to a lesser extent, male White perch tend to venture along the southwestern shore of Île d'Orléans; that behavior is no longer apparent after spawning. Female White perch have a broader distribution after spawning than did male White perch. With the notable exception of female White perch, the model predicted a broader distribution of fish inside and outside the study area during than after spawning. Hence, female White perch appear prevalent in the middle of CIO south and north of the road bridge after spawning. Male White perch do not seem to modulate their distribution, with respect to spawning timing, as deeply as did female White perch.

Discussion

In the present study, we described the construction of an ANN-based NHM in estuarine conditions and applied it to predict fish probability of presence in a planned road bridge construction area. The NHM we obtained has a broad scope,

encompassing different sex, species, and time of the year. We found that ANN model to have good performance and we expect it to be useful for assessing the impact of potential alterations of Lake sturgeon and White perch habitat during and after construction works. These works will both temporarily and permanently alter substrate composition, depths, and local hydraulics in various ways. To assess the effect of a given construction scenario, one would input its associated transient hydrological alterations, combined them with tidal predictions for the scheduled construction period, simulate hydraulics, and finally assess changes in fish distribution using the NHM. Assessing permanent alterations would involve a similar procedure while using average yearly tidal predictions instead of time-specific ones.

In cross-validation trials, ANN-based NHM surpassed a logistic regression-based NHM built in a similar fashion by a correct classification margin of 6.0%; a margin that may seem small but becomes increasingly difficult to improve as it approaches perfection (i.e., 100% correct classification). A 6.0% improvement over a model having an out-of-the-sample performance of 72.5% thus represents a relative improvement of $\frac{6.0}{100-72.5} \approx 21.6\%$ over the remaining margin for improvement. That relative improvement appears slightly greater ($\frac{6.5}{100-70.8} \approx 22.3\%$) when assessed over the training data, but shrinks somehow ($\frac{4.9}{100-64.8} \approx 14.0\%$) when assessed on the testing data. An other highlight of the present study is ANN's ability for implicitly representing relationships with a vast array of different shapes, and involved in an equally vast array of possible interaction patterns, without them having to be explicitly enumerated by the model builder. Individuals distribute differently in the study area as a function of their species and genders.

The actual road bridge is located well into the habitat used by both species during (and, for Lake sturgeon, also after) their reproductive period. We therefore expect that any construction work undertaken in the bridge area could add an impact over that already present from other past and present human activities. However, that impact need not be negative. On the habitat quality maps (Fig. 6, 7; *Supplementary material*, figs. 20-91), the areas in the direct vicinity of the actual road bridge

(< 100 m) generally appear to have a higher quality than their surroundings located farther away (> 100 m). That difference is readily visible for both species during the spawning period (Fig. 6; even more apparent at high tide, see *Supplementary material*, Figs. 43-45). Since the actual road bridge now has and apparently positive local effect on habitat quality, its possible destruction following the construction of a replacement structure might have a detrimental effect. We are hopeful that the ANN model we developed in the present study will be helpful in planning future development work in that area, from the construction of new infrastructures (e.g., bridge, split road), the reuse or disposal of existing ones, to the planning of any compensatory measures that may be needed.

Fish tracking information was obtained for only a part of the year, from late April to mid-November. In addition to the increasing costs and logistic complexity, year-round fish tracking would also have carried its share of uncertainties as the harsh conditions found in the area involve shifting winter ices and spring flooding that can displace the VR2W hydro-acoustic receivers off their operational location or wash them away from the surveying area. However, hydro acoustic hyperbolic ranging operates well under winter conditions, provided that receivers can remain at a steady location. One way to achieve overwinter surveying may involve, for instance, technologies for stowing receivers near the bottom or within the sediments whenever water becomes too shallow and ices or floodwaters threaten receivers. Although it is computationally possible to extrapolate the model beyond the period of time for which data is available, that practice would yield questionable results about the distribution of the species. Extrapolating any of the descriptors beyond the range that has been observed during the study period would pose a similar issue: there would be no way to ascertain that the shape of the relationships described by the model would still follow the trends occurring in the field. Hence, the ability of ANN models to represent complex, non-linear, patterns also make them potentially more vulnerable to extrapolation then the more classical regression models, as the former's response may experience steeper gradients on descriptors extremes (or for oddly-observed combinations thereof). The construction of a large infrastructure

such as a road bridge spans many years and the effect of the structure will last, year-round, for decades to follow. Therefore, it may ultimately be worthwhile to address the question of year-round fish distribution, in future studies, by extending fish sampling effort to parts of the year that have not yet been sampled.

Pseudo absence

A noteworthy issue with telemetry studies, which we mentioned earlier, is that it does not readily identify the conditions that are not sought after (or avoided) by the tracked organisms. Absence data have to be drawn in some way. Here we performed that drawings on pools of locations chosen to feature conditions that were not too similar to those prevailing on observation sites. From a statistical perspective, such a decision with regard to potential PA pooling assumes that the null hypothesis, whereby fish move about the experimental area irrespective of the conditions found therein, is rejected first hand. Putting all sites in the pool, irrespective of conditions found in them, would have come with the initial assumption that the null hypothesis is true. Hence, PA drawing, although arguably a widespread practice, is somewhat remote from the ideal as there is no way to correctly draw absence data without making an *a priori* statement about whether or not the environment influences distribution. Drawing locations randomly assumes a null distribution of the organisms (i.e., one that is purely random, unrelated to environmental conditions) for the locations marked as absences. However, locations marked as presence will not conform with a null distribution under scenarios where distribution can be modeled. Assuming that the distribution is indeed influenced by the environment (i.e., isn't null), a model that is consequent with the alternate hypothesis (i.e., that distribution is related with the environment) will be obtained by discarding locations that are similar to those where organisms were observed. In that scenario, observing an organism at some places can be regarded as an evidence that conditions found therein are prized by the organism being tracked.

As it turns out, it is impossible to generate PA data that are both statistically unbiased in terms of α (i.e., the probability to reject the hypothesis of a null

distribution when distribution is actually unrelated to the environment) and β (i.e., the probability to accept the hypothesis of a null distribution when distribution is actually related to the environment) error rates without knowing the outcome of the underlying hypothesis test of the effect of the environment on distribution first hand, thereby making that whole decision process circular by nature. Here, we weren't primarily concerned with hypothesis testing; by drawing PA in a stratified fashion (i.e., by removing locations with conditions similar to those of observation sites from the pool), we thus assumed fish distribution to be related to the environment. It is worthwhile to remain mindful of the fact that our apparent success at correctly predicting PA to a greater extent than expected by chance alone (e.g., the correct classification rate of 60.6% obtained for the deep ANN NHM for cross-validation trials whereas 50% would be expected) was likely a consequence of beginning from that assumption. Another trade-off to obtain the most reliable distribution model in such an arguably imperfect framework could have been to draw PA in an entirely random fashion and select hyper-parameters to obtain a model maximizing the correct classification of presence data while keeping the correct classification rate for PA as close as possible to their expected null value (50%).

Beyond a binary distribution model

For the future, we envision another modeling framework, besides the binary classification framework used in the present study, that does away with the need to draw PA and associated shortcomings (i.e., a so-called "presence-only" model). Also, that framework would make use of the information related to the movements of organisms in time, which is currently being overlooked by binary classification models. ANN being so versatile and adaptable, they can be used to predict transition among discrete states within a set of possible states; not unlike a ANN-based equivalent of a multi-state (linear) Markov model. Such a model would operate on a discretized map of homogeneous habitat polygons and predict the conditional probability, given the prevailing environmental conditions, for staying in or moving about polygons after a given amount of time. These polygons of

homogeneous environmental conditions can be readily obtained using a spatially-constrained cluster analysis (Legendre and Legendre, 2012).

There may be several ways by which such an habitat transition model could be implemented. Here, we propose the following workflow. Firstly, one defines a naive transition classifier ANN representing the organism's marginal probability of distribution among the different habitat states (presence in a polygon). That model thus takes organisms previous locations (i.e., a set binary descriptors), together with the amount of time elapsed since that previous observation (i.e., a single continuous descriptor) as its input layer, process it through a few hidden layers, and ends into a softmax (i.e., multiple outcomes) output layer predicting the probability of presence of the fish in the different habitat polygons. Trained strictly on the output data, that classifier plays a role similar to that of the exponential of the time-multiplied transition intensity matrix in a multistate Markov (chain) model. Secondly, a variational auto-encoder VAE is built on the environmental conditions data. A VAE is a non-supervised ANN that re-generate its own set of input data while forcing information through a purposely-made information bottleneck in the form of a latent vector space constrained to have a user-defined statistical distribution. It does so by first encoding the information content of the variables into a compressed representation, with specified statistical properties, using a convergent network called the encoder and then turning this representation back into the original representation using a divergent network called the decoder. The encoder network thus obtained can be recycled into another model and provide it with a simplified representation of the input variables. Although that step is optional, we anticipate that a preliminary processing of the environmental variable space into a latent vector space would later simplify the injection of environmental information into the classifier. Thirdly, encoding section of the VAE developed in the second step is connected to the classifier. That connection would operates through sets of weights injecting the latent vector representation of the environment into some of the classifier's hidden layers, thereby turning the whole network into a conditional (non-naive) transition classifier. The latter connecting weights would be estimated

while keeping that pertaining strictly to the naive transition classifier and environmental VAE sections constant during the back-propagation estimation process. Optionally, the resulting compound model might be trained for some more epochs while allowing all weights to be updated for refinement purposes. Also, we may use two copies of the same VAE (i.e., sharing weights and providing the same latent vector representation), one for describing conditions at previous time step, and the other describing the actual conditions experienced by the organism.

Conclusion

In spite of a few limitations related to sampling, we have shown here that the most recent tools in acoustic positioning, remote sensing, tidal and hydraulics modeling, and machine learning can be combined to build habitat probabilistic models for estuarine environments. The resulting model has been found to be powerful within its seasonal scope. It is our dearest hope that the results featured in the present study inspire other multi-disciplinary teams to contribute further advancements in compounding the use of advanced remote sensing, 2D (or 3D) physical environment modeling, large-scale field tracking, and deep learning analysis methods for habitat modeling. Indeed, there are growing needs for characterizing the habitat requirements for the ever-growing list of organisms that are affected by human activity. Safeguarding these organisms and the various ecosystems that they inhabit gives us many reasons to look forward to the deployment of integrative habitat modeling efforts.

Acknowledgments

We are immensely thankful to our dear colleagues who helped us with the various methodological and programmatic aspects of our study. We would also like to thank the many generous people how share their knowledge freely and openly in the various code library, tutorials, forums, science blogs, etc., found nowadays on the

Internet. They allowed us (along with many others) to obtain timely answers to the myriads of small (and sometimes bigger) questions that arose incessantly during our work. It is fair to say that we would have had a much harder time without them.

Figure legend

Figure 1: Map of section of the St. Lawrence Fluvial Estuary (SLFE) where hydraulic modeling was performed (i.e., modeled area), which includes the fish habitat study area (shown in the upper left hand inset). The legend in the upper portion of the figure contains the colors used to identify the different terrain features, whereas that in the lower right end references the place names mentioned in the text with numbers. The fine dotted curves indicate reference latitudes (Lat) and longitudes (Lon).

Figure 2: Estimation of the first derivative of a tidal height time series with respect to time. The time series is sampled at a suite of times (dots) located before and after the reference time $t = 0$ (here, 13:30:00). From these sampling points, a locally-defined third order polynomial curve is fitted and its tangent at the reference time (obtained from the derivative of the polynomial) is taken as the estimate of the time derivative of the time series at that particular moment. The procedure is repeated for every sampling times, yielding a time series that is the numerical first derivative of the original time series (tidal rate).

Figure 3: Example of the temporal variable pair used to represent the seasons in the numerical habitat model (Spr: spring, Sum: summer, Aut: autumn, Win: winter) in the model.

Figure 4: Example of the temporal variable pair used to represent the time of day in the numerical habitat model.

Figure 5: Probability distribution of the fish presence model response when fish were present (solid, tracked using acoustic telemetry) and for pseudo absence (dashed) drawn in the study area, with the model's decision threshold for absence ($Pr < 0.4143$) and presence ($Pr \geqslant 0.4143$). The model is a three-layer ANN.

Figure 6: Map of the fish habitat during the spawning period (May 01 to June 30) obtained for female and male Lake sturgeon (*A. fulvescens*) and White perch (*M. americana*) from the deep ANN fish habitat model developed in the present study. Shades of gray represent the proportion of the time the model predicted fish to be present on locations that were was within the depth range of the model ($[0.3 - 30\,\mathrm{m}]$) (see the legend in the upper part of Fig. 1 for the meaning of the other colors on the map).

Figure 7: Map of the fish habitat after the spawning period ended (July 01 to Nov. 21) obtained for female and male Lake sturgeon (*A. fulvescens*) and White perch (*M. americana*) from the deep ANN fish habitat model developed in the present study. Shades of gray represent the proportion of the time the model predicted fish to be present on locations that were was within the depth range of the model ($[0.3 - 30\,\mathrm{m}]$) (see the legend in the upper part of Fig. 1 for the meaning of the other colors on the map).

References

Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Meth. Ecol. Evol.*, 3:327–338.

Bivand, R., Keitt, T., and Rowlingson, B. (2017). *rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 1.2-8.

Boisclair, D. (2001). Fish habitat modeling: from conceptual framework to functional tools. *Can. J. Fish. Aquat. Sci.*, 58:1–9.

Brosse, S., Guegan, J.-F., Tourenq, J.-N., and Lek, S. (1999). The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. *Ecol. Model.*, 120:299–311.

Candel, A., LeDell, E., Parmar, V., and Arora, A. (2018). Deep learning with H2O, January 2018: sixth edition.

687 Capra, H., Plichard, L., Bergé, J., Pella, H., Ovidio, M., McNeil, E., and
688 Lamouroux, N. (2017). Fish habitat selection in a large hydropeaking river:
689 Strong individual and temporal variations revealed by telemetry. *Sci. Total*
690 *Environ.*, 578:109–120.

691 COSEWIC (2017). Assessment and status report on the lake sturgeon *Acipenser*
692 *fulvescens*, western hudson bay populations, saskatchewan-nelson river
693 populations, southern hudson bay-james bay populations, great lakes-upper st.
694 lawrence populations in canada.

695 Dumas, F. and Langlois, G. (2009). MARS, Model for Application at Regional
696 Scale, Scientific model description. Technical report, Ifremer.

697 Foubert, A., Pichon, C. L., Mingelbier, M., Farrell, J. M., Morin, J., and Lecomte,
698 F. (2019). Modeling the effective spawning and nursery habitats of northern pike
699 within a large spatiotemporally variable river landscape (St. Lawrence River,
700 Canada). *Limnology and Oceanography*, 64(2):803–819.

701 Gibson, R. N. (1993). Intertidal teleosts: life in a fluctuating environment. In
702 Pitcher, T. J., editor, *Behaviour of Teleost Fishes*, pages 513–536. Chapman and
703 Hall, London, U. K., second edition.

704 Guay, J. C., Boisclair, D., Rioux, D., Leclerc, M., Lapointe, M., and Legendre, P.
705 (2000). Development and validation of numerical habitat models for juveniles of
706 atlantic salmon (*Salmo salar*). *Can. J. Fish. Aquat. Sci.*, 57:2065–2075.

707 Guénard, G., Lanthier, G., Harvey-Lavoie, S., Macnaughton, C. J., Senay, C.,
708 Lapointe, M., Legendre, P., and Boisclair, D. (2017). Modelling habitat
709 distributions for multiple species using phylogenetics. *Ecography*, 40(9):1088–1097.

710 Hastie, T. J. and Pregibon, D. (1991). *Generalized linear models*, volume Statistical
711 models in S, chapter 6, pages 195–247. Wadsworth, Pacific Grove, CA.

712 Heniche, M., Secretan, Y., Boudreau, P., and Leclerc, M. (2006). A two-dimensional

713 finite element drying-wetting shallow water model for rivers and estuaries. *Adv.*
714 *Water Resour.*, 23:359–372.

715 Hijmans, R. J. (2016). *raster: Geographic Data Analysis and Modeling*. R package
716 version 2.5-8.

717 Legendre, P. and Legendre, L. (2012). *Numerical Ecology, Third English Edition.*
718 Elsevier Science B.V., Amsterdam, The Netherlands.

719 Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., and Aulagnier, S.
720 (1996). Application of neural networks to modelling nonlinear relationships in
721 ecology. *Ecol. Model.*, 90:39–52.

722 Lipton, Z. C., Elkan, C., and Naryanaswamy, B. (2014). Thresholding classifiers to
723 maximize f1 score. *arXiv stat.ML*, 1402.1892v2.

724 Lotze, H. K., Lenihan, H. S., Bourque, B. J. Bradbury, R. H., Cooke, R. J., Kay,
725 M. C., Kidwell, S. M., Kirby, M. X. Peterson, C. H., and Jackson, J. B. C. (2006).
726 Depletion, degradation, and recovery potential for estuaries and coastal seas.
727 *Science*, 312:1806–1809.

728 Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proc. Nat.*
729 *Instit. Sci. India*, 2(1):49–55.

730 Matte, P., Jay, D. A., and Zaron, E. D. (2013). Adaptation of classical tidal
731 harmonic analysis to nonstationary tides, with application to river tides. *J.*
732 *Atmos. Oceanic Technol.*, 30(3):569–589.

733 Matte, P., Secretan, Y., and Morin, J. (2014a). Quantifying lateral and intratidal
734 variability in water level and velocity in a tide-dominated river using combined
735 RTK GPS and ADCP measurements. *Limnol. Oceanogr. Methods*, 12:281–302.

736 Matte, P., Secretan, Y., and Morin, J. (2014b). A robust estimation method for
737 correcting dynamic draft error in PPK GPS elevation using ADCP tilt data. *J*
738 *Atmos. Oceanic Techol.*, 31:729–738.

739 Matte, P., Secretan, Y., and Morin, J. (2014c). Temporal and spatial variability of
740 tidal-fluvial dynamics in the St. Lawrence fluvial estuary: An application of
741 nonstationary tidal harmonic analysis. *J. Geophys. Res.*, 119:5724–5744.

742 Matte, P., Secretan, Y., and Morin, J. (2017a). Hydrodynamic modeling of the st.
743 lawrence fluvial estuary. i: Model setup, calibration, and validation. *J. Waterway,*
744 *Port, Coastal, Ocean Eng.*, 143:04017010.

745 Matte, P., Secretan, Y., and Morin, J. (2017b). Hydrodynamic modeling of the st.
746 lawrence fluvial estuary. ii: Reproduction of spatial and temporal patterns. *J.*
747 *Waterway, Port, Coastal, Ocean Eng.*, 143:04017011.

748 Mingelbier, M., Brodeur, P., and Morin, J. (2008). Spatially explicit model
749 predicting the spawning habitat and early stage mortality of northern pike (esox
750 lucius) in a large system: the st. lawrence river between 1960 and 2000.
751 *Hydrobiologia*, 601:55–69.

752 Morin, J., Bachand, M., Richard, J. H., and Martin, S. (2016). Modeling the rainy
753 lake and namakan reservoir ecosystem response to water level regulation.
754 Scientific Report SR-110, Hydrology and Ecohydraulic Section, Environment
755 Canada, Québec. Prepared for the International Joint Commission study on
756 Rainy Lake and Namakan Reservoir.

757 Morin, J., Boudreau, P., Secretan, Y., and Leclerc, M. (2000). Pristine lake
758 saint-françois, st. lawrence river: Hydrodynamic simulation and cumulative
759 impact. *J. Great Lakes Res.*, 26(4):384–401.

760 Morin, J., Mingelbier, M., Bechara, J., Champoux, O., Y, S., Jean, M., and
761 Frenette, J.-J. (2003). Emergence of new explanatory variable for 2d habitat
762 modelling in large rivers: the St. Lawrence experience. *Can. Water Res. J.*,
763 28:249–272.

764 Mukhin, D., James, D. A., and Luciani, J. (2016). *OCI Based Oracle Database*
765 *Interface for R.* R package version 1.3-1.

766 Olden, J. D. and Jackson, D. A. (2001). Fish-habitat relationships in lakes: Gaining
767 predictive and explanatory insight by using artificial neural networks. *Trans. Am.*
768 *Fish. Soc.*, 130:878–897.

769 Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in r.
770 *R News*, 5:9–13.

771 Quetglas, A., Ordines, F., and Guijarro, B. (2011). *The Use of Artificial Neural*
772 *Networks (ANNs) in Aquatic Ecology.*, chapter Artificial Neural Networks -
773 Application, chapter 27, pages 567–586. InTech.

774 R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R
775 Foundation for Statistical Computing, Vienna, Austria.

776 R Special Interest Group on Databases (R-SIG-DB), Wickham, H., and Müller, K.
777 (2016). *DBI: R Database Interface*. R package version 0.5.

778 Roessig, J., Woodley, C., Cech, J. J. J., and Hansen, L. (2004). Effects of global
779 climate change on marine and estuarine fishes and fisheries. *Rev. Fish Biol.*
780 *Fisheries*, 14:251–275.

781 Sagarese, S. R., Frisk, M. G., Miller, T. J., Sosebee, K. A., Musick, J. A., and Rago,
782 P. J. (2014). Influence of environmental, spatial, and ontogenetic variables on
783 habitat selection and management of spiny dogfish in the Northeast (US) shelf
784 large marine ecosystem. *Can. J. Fish. Aquat. Sci.*, (71):567–580.

785 Skov, H., Humphreys, E., Garthe, S., Geitner, K., Gremillet, D., Hamer, K. C.,
786 Hennicke, J., Parner, H., and Wanless, S. (2008). Application of habitat
787 suitability modelling to tracking data of marine animals as a means of analyzing
788 their feeding habitats. *Ecol. Model.*, 212(3-4):504–512.

789 Spruzen, F. L., Richardson, A. M. M., and Woehler, E. J. (2008). Influence of
790 environmental and prey variables on low tide shorebird habitat use within the
791 robbins passage wetlands, northwest tasmania. *Estuarine Coastal Shelf Sci.*,
792 78(1):122–134.

The H2O.ai team (2017). *h2o: R Interface for H2O*. R package version 3.10.5.3.

Valiquette, E., Legault, M., and Harvey, V. (2016). État référence de la faune aquatique et de ses habitats dans le secteur du pont de l'île d'orléans: rapport final. première partie — description physique et inventaires biologiques. Technical report, Ministère des Forêts, de la Faune et des Parcs, Secteur Faune et Parcs, Direction générale de la gestion de la faune et des habitats, Direction de l'expertise sur la faune aquatique, Québec, xxviii + 199 p.

VanDerWal, J., Shoo, L. P., Graham, C., and Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecol. Model.*, 220(4):589–594.

Wickham, H., James, D. A., and Falcon, S. (2014). *RSQLite: SQLite Interface for R*. R package version 1.0.0.

Figures

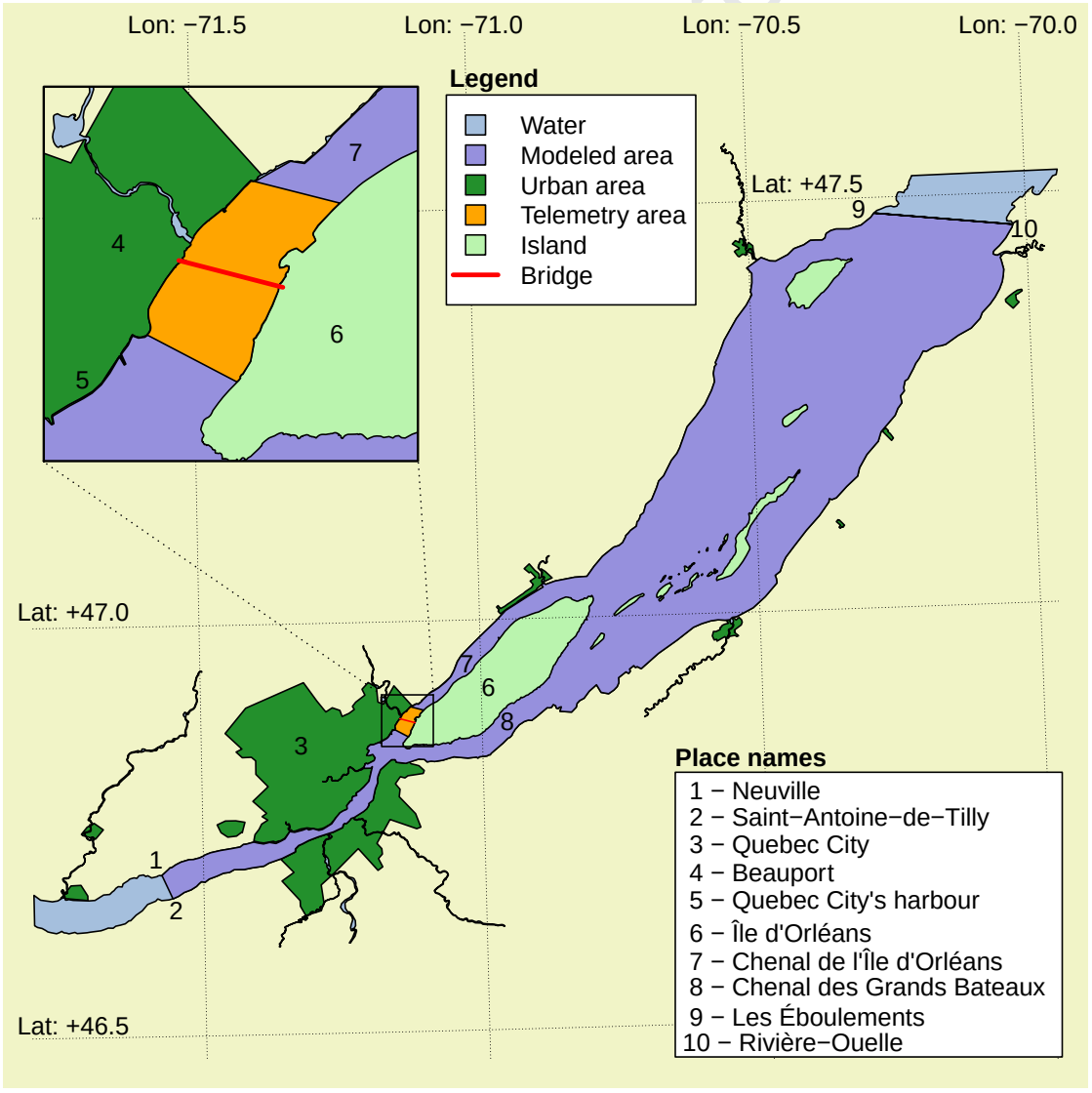
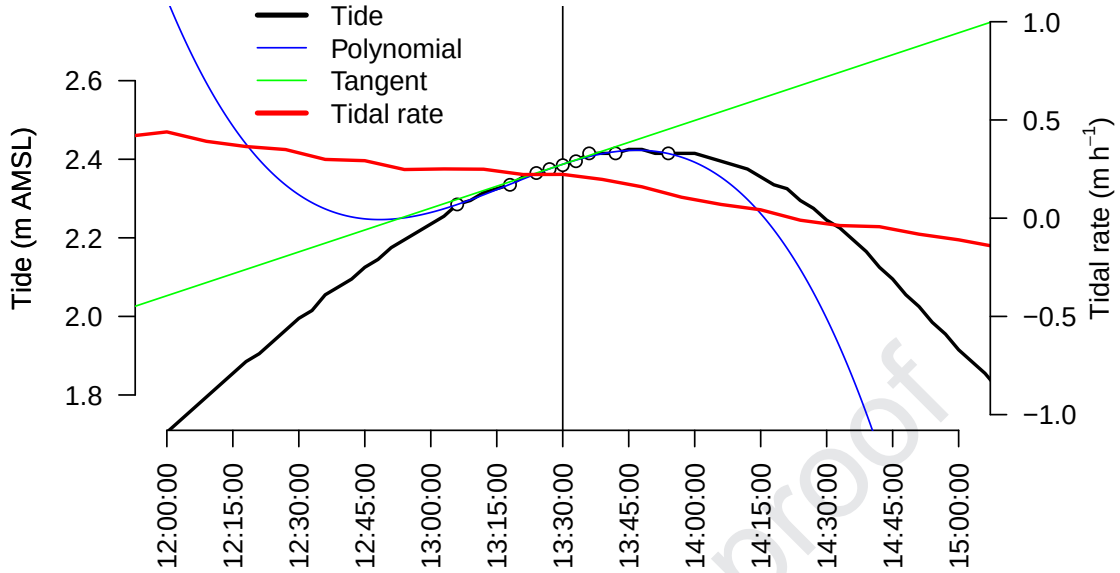
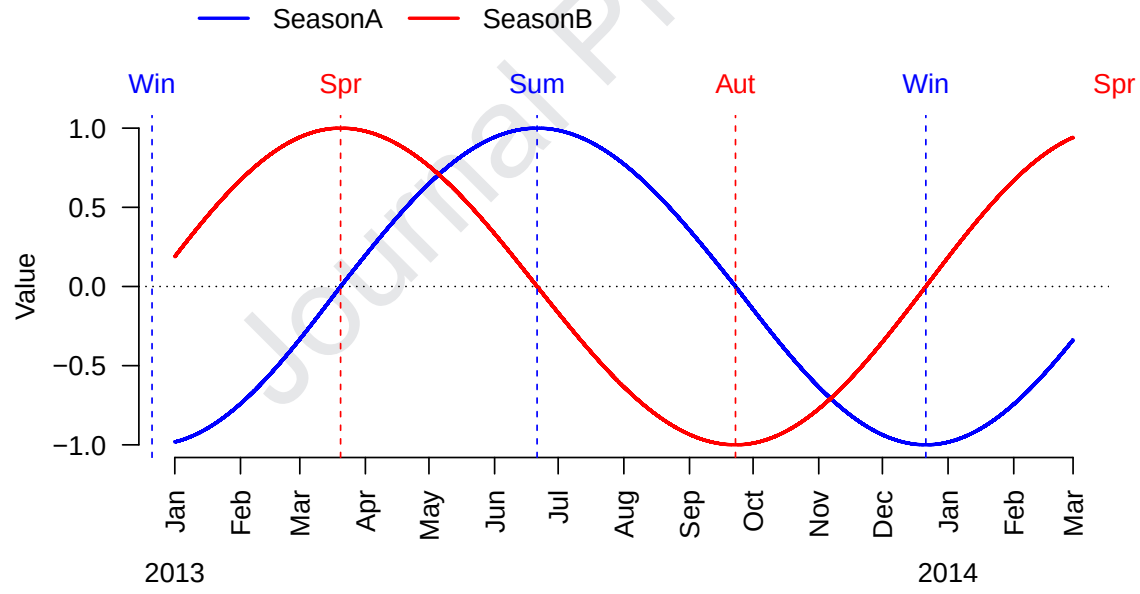


Figure 1



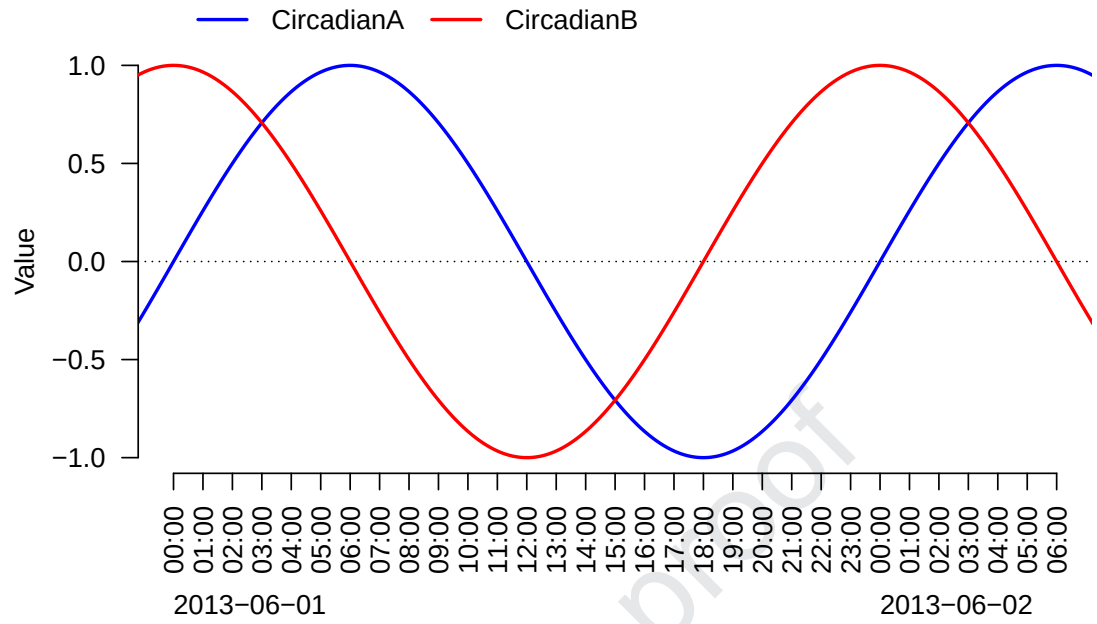
808

809 Figure 2



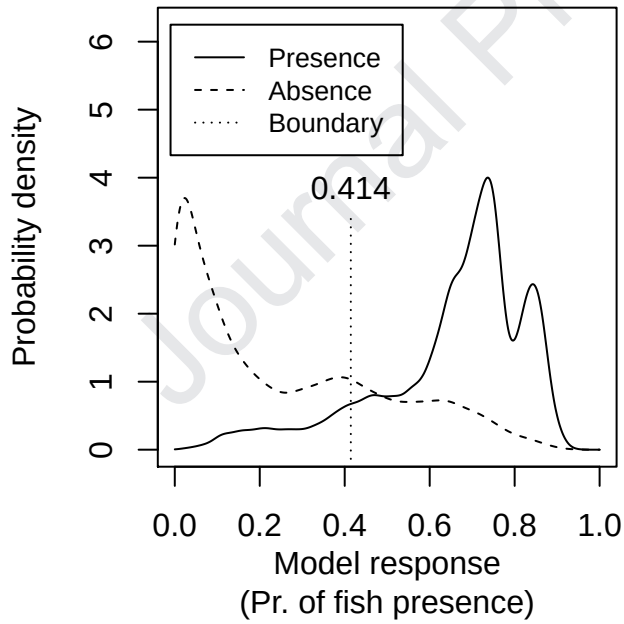
810

811 Figure 3



812

813 Figure 4



814

815 Figure 5

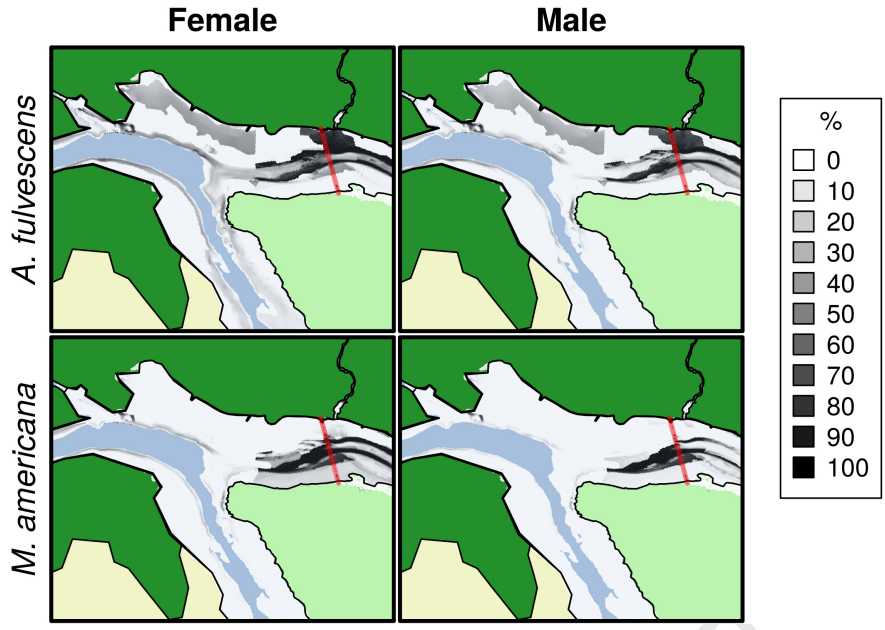


Figure 6

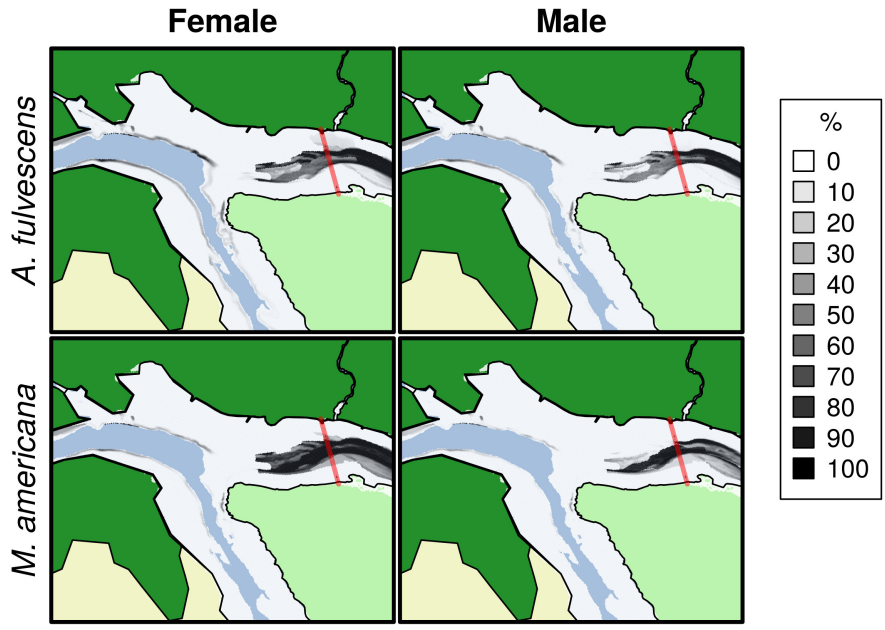


Figure 7

Highlights for paper "Deep learning habitat modeling for moving organisms in rapidly changing environments: an example of two fishes in a tidal estuary"

Guillaume Guénard, Jean Morin, Pascal Matte, Yves Secretan, Sylvain Martin, Eliane Valiquette, and Marc Mingelbier

6 Feb. 2020

Highlights

- We modeled the habitat of two fishes: Lake sturgeon and White perch
- Habitat is located in the St. Lawrence River estuary, near Île d'Orléans
- We used a deep Artificial Neural Network model with many types of descriptors
- The model classified 94.0% of observations as presence during cross-validation
- Descriptors often displayed non-linearity and varied among the environment

Environment and
Climate Change Canada

Environnement et
Changement climatique Canada

2020-02-06

Editorial office, Estuarine, Coastal, and Shelf
Science

Review of manuscript ***"Deep learning habitat modeling for moving organisms in rapidly changing estuarine environments: a case of two fishes"*** **(Research article)**

Greetings,

We hereby declare that none of us authors have conflicts of interest of any sort related to the content and publishing of the present study.

Yours sincerely,
DR. GUILLAUME GUÉNARD - Modeling Project Manager
National Hydrological Service
Environment and Climate Change Canada
1550 avenue d'Estimauville,
Quebec City QC, G1J 5E9 Canada