*North American Journal of Fisheries Management*

ISSN: 0275-5947 print / 1548-8675 online
DOI: 10.1002/nafm.10380

ARTICLE

# Morphological Identification of Bighead Carp, Silver Carp, and Grass Carp Eggs Using Random Forests Machine Learning Classification

**Carlos A. Camacho,*[1] Christopher J. Sullivan, and Michael J. Weber**
*Department of Natural Resource Ecology and Management, Iowa State University, 339 Science Hall II, Ames, Iowa 50011, USA*

**Clay L. Pierce**
*U.S. Geological Survey, Iowa Cooperative Fish and Wildlife Research Unit, Ames, Iowa 50011, USA*

***Abstract***

**Visual identification of fish eggs is difficult and unreliable due to a lack of information on the morphological egg characteristics of many species. We used random forests machine learning to predict the identity of genetically identified Bighead Carp *Hypophthalmichthys nobilis*, Grass Carp *Ctenopharyngodon idella*, and Silver Carp *H. molitrix* eggs based on egg morphometric and environmental characteristics. Family, genus, and species taxonomic-level random forests models were explored to assess the performance and accuracy of the predictor variables. The egg characteristics of Bighead Carp, Grass Carp, and Silver Carp were similar, and they were difficult to distinguish from one another. When combined into a single invasive carp class, the random forests models were ≥ 97% accurate at identifying invasive carp eggs, with a ≤5% false positive rate. Egg membrane diameter was the most important predictive variable, but the addition of ten other variables resulted in a 98% success rate for identifying invasive carp eggs from 26 other upper Mississippi River basin species. Our results revealed that a combination of morphometric and environmental measurements can be used to identify invasive carp eggs. Similar machine learning approaches could be used to identify the eggs of other fishes. These results will help managers more easily and quickly assess invasive carp reproduction.**

Aquatic nuisance species are becoming more common in the United States, and they are expanding their distribution through both natural and anthropogenic dispersal (Lodge 1993; Rahel 2002; Kolar et al. 2007). First introduced in the 1960s, Grass Carp *Ctenopharyngodon idella*, Silver Carp *Hypophthalmichthys molitrix*, and Bighead Carp *H. nobilis*, collectively called "invasive carp" hereafter, have invaded the Mississippi River basin and they are expanding their range, threatening ecosystem integrity (Freeze and Henderson 1982; Wittmann et al. 2014). Efforts to determine areas of current and potential establishment have largely relied upon the detection of early life stages (Deters et al. 2013; Coulter et al. 2016; Embke et al. 2016). However, discrepancies and a lack of information that describes the morphological egg characteristics of invasive carp and native North American species has made the visual identification of fish eggs difficult and unreliable (Richards 1985; Larson et al. 2016). To avoid the inconsistencies of visual identification, genetic analysis is often the preferred method for egg identification (Becker et al. 2015; Coulter et al. 2016; Embke et al. 2016). Unfortunately, genetic analysis is expensive, making it impractical for use on the large quantities of eggs that are commonly captured during ichthyoplankton sampling.

*Corresponding author: ccamacho0526@hotmail.com
[1]Present address: Idaho Department of Fish and Game, 2885 West Kathleen Avenue, Coeur d'Alene, Idaho 83815, USA.
Received May 7, 2019; accepted October 11, 2019

Therefore, cost effective egg identification techniques must be developed to ensure the accurate and timely detection of the establishment of invasive carp from large sample collections, which is vital to rapid response efforts.

Invasive carp eggs have been described in great detail (Chapman and George 2011; George and Chapman 2013, 2015), but similar detailed visual descriptions do not exist for many native upper Mississippi River fish species. Consequently, fish eggs that were collected in pools 9 and 11 of the upper Mississippi River were falsely classified as invasive carp based on membrane size but were later genetically identified as a native cyprinid (Larson et al. 2016). The misclassified eggs had a larger membrane diameter than did any native fish eggs that had been previously reported in literature. Insufficient knowledge about the natural variation of egg morphology within a species due to biotic (e.g., female size and fitness; Crean and Marshall 2009) and abiotic (e.g., water temperature; Hutchings 1991) factors can result in inaccurate identification protocols. Furthermore, the egg morphology of invasive carp may be different or display greater variability in newly invaded systems due to the wide range of environmental variability and lack of natural stressors in nonnative systems compared with their native range (Mack et al. 2000; Peterson and Vieglais 2001; Lenaerts et al. 2015). Thus, additional information on fish egg morphology among native and nonnative species is needed to clarify and refine the distinguishing visual characteristics that are necessary for correctly identifying eggs in the Mississippi River basin.

The choice of preservative is of great importance to morphometric analysis due to differential physical changes of samples following their preservation (Martinez et al. 2012). Measurements from live specimens offer the best morphological descriptions of natural conditions. However, obtaining accurate field measurements of live eggs at the time of capture is usually not possible and preservation is required for storage and transport (Kelso and Rutherford 1996). A growing body of literature suggests that all forms and combinations of preservation and fixation change the morphology (e.g., shape, size, and weight) of the eggs (Kelso and Rutherford 1996; Frimpong and Henebry 2012), including commonly used preservatives such as formalin (König and Borcherding 2012) and ethanol (Martinez et al. 2012). Unfortunately, most egg descriptions in the literature, including those for invasive carp, are of live specimens (Yi et al. 2006; George and Chapman 2013, 2015) and they do not translate well to preserved specimens (Martinez et al. 2012). Therefore, the existing body of literature that describes egg morphology is only applicable to a small subset of studies.

The objective of this study was to assess whether random forests machine learning could be used to accurately identify invasive carp eggs that have been preserved in ethanol. Formalin is often the preferred preservative for ichthyoplankton specimens (Kelso and Rutherford 1996). However, formalin degrades DNA quality through the preservation process, hindering genetic identification (Wiegand et al. 1996; Diaz-Viloria et al. 2005). In contrast, ethanol preservation does not affect the integrity of DNA and is preferred over formalin for material that is subjected to molecular techniques. We used random forests machine learning to predict the classification of genetically identified eggs based on morphological and environmental characteristics. First, we examined several random forests models to determine the taxonomic resolution that was best suited to accurately predicting invasive carp. Second, we combined all of the invasive carp species into a single group and reexamined the best taxonomic resolution for predicting invasive carp. Third, we used a variable importance measure to determine which variables most accurately predicted invasive carp eggs. The results of this project provide a quantitative tool that provides a cost-effective technique for detecting invasive carp reproduction.

## METHODS

Fish eggs were sampled across 2 years at nine locations along the northern edge of invasive carp reproduction within the upper Mississippi River and the lower portions of four major tributaries in southeast Iowa (Figure 1). At each location, a single transect was established consisting of a straight line from streambank to streambank, perpendicular to the main flow of the water. Three sample sites were located on each transect in the thalweg, channel border, and backwater habitats. Egg sampling was conducted at each habitat in each transect every 10 d from late April through the end of September in 2014 and 2015. The habitats were defined by the magnitude of water flow. The thalweg habitat was located in the portion of the river with the fastest flowing water, typically in the main channel. The backwater habitat consisted of areas with little or no flow such as inside river bends, sloughs, and inundated floodplains. The channel border habitat had an intermediate flow relative to the thalweg and backwater habitats within the same transect. Eggs were collected with an ichthyoplankton net (0.5 m diameter opening, 500-μm square mesh) that was towed adjacent to the boat in an upstream direction just below the water surface for a maximum of 4 min (depending on debris load). The boat speed was kept at a constant relative to the shoreline or at boat motor idle if the river flow was minimal. During each tow, water temperature (°C) and conductivity (μS) was measured with an ExtStik II Conductivity Meter (Extech Instruments Corporation, Nashua, New Hampshire). After each tow, the contents of the net were washed into a collection cup on the cod end, drained of water, placed into jars, and preserved with 95% ethanol.

In the laboratory, the eggs were separated from the debris by at least two individuals until no additional eggs were found. The eggs from each tow were stored in 20-mL glass scintillation vials with 95% ethanol for no longer than 6 months before being photographed. More eggs were collected than could be genotyped. Therefore, a random subsampling scheme representative of each tow was used to capture any spatiotemporal variation within the species assemblages and egg morphology variation within a single species (Hutchings 1991). Each subsampled egg was photographed (with an Olympus SZX7 microscope; Image Pro 7.0 software, Media Cybernetics, Bethesda, Maryland) at 2× magnification in a petri dish with just enough ethanol to cover the egg and to aid in holding the egg stationary. For eggs with an embryo, the pictures were taken in the dorsal, ventral, and lateral positions in relation to the embryo. If an embryo was not identifiable, a picture was taken after a quarter rotation of the egg on its *y*-axis, *x*-axis, and again on its *y*-axis. After photographing, each egg was stored individually in a 5-mL microcentrifuge tube with 95% ethanol for genetic analysis.

From the pictures, the eggs were visually categorized and measured by using Image Pro software. The eggs were first identified as either having a definable embryo element within the membrane or not having any discernable embryo (Figure 2). The embryos were further classified based on the egg development stages defined in Kelso and Rutherford (1996). All of the embryos were examined for the presence or absence of pigment (Figure 2). The egg membranes were further classified based on the presence or absence of a deflated membrane and debris adhesion. Deflated membranes were characterized by a nonspherical shape with wrinkles, and debris adhesion was characterized by organic or inorganic material sticking to the egg membrane (Figure 2). As defined by Kelso and Rutherford (1996), four diameter measurements (mm) with starting points that were equally spaced around the circumference were taken from the membrane and embryo as well as a total length measurement (mm) along the midline from all of the late stage embryos (Figure 3). The average, standard deviation, and coefficient of variation of the membrane and embryo diameters for each egg were calculated. Lastly, the visually transparent region between the embryo and outer membrane, known as the perivitelline space, was calculated as follows:

$$\text{Perivitelline space index} = 1 - \left( \frac{\text{embryo average}}{\text{membrane average}} \right).$$

DNA was extracted from the individual eggs by using the Gentra Puregene Tissue Kit (Qiagen, Germantown, Maryland) or the Promega Wizard Genomic DNA Purification kit (Promega Corp., Madison, Wisconsin) according to the manufacturer's suggested protocol and stored at

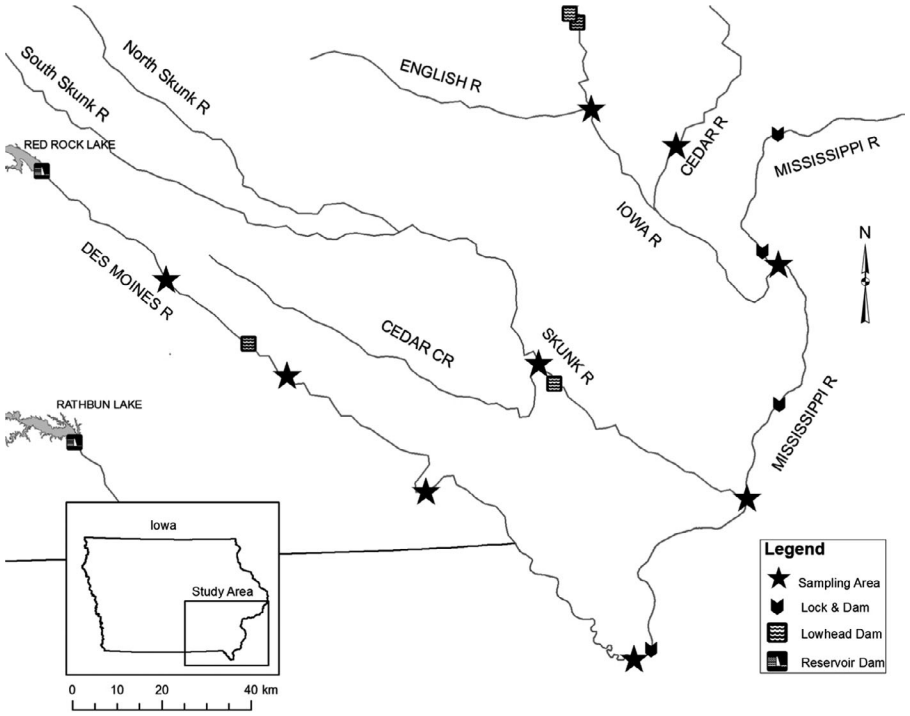


FIGURE 1. Approximate sampling sites, fish passable lowhead dams, fish barrier reservoir dams, and lock and dams locations in the Des Moines, Skunk, Iowa, Cedar and upper Mississippi rivers across southeastern Iowa.

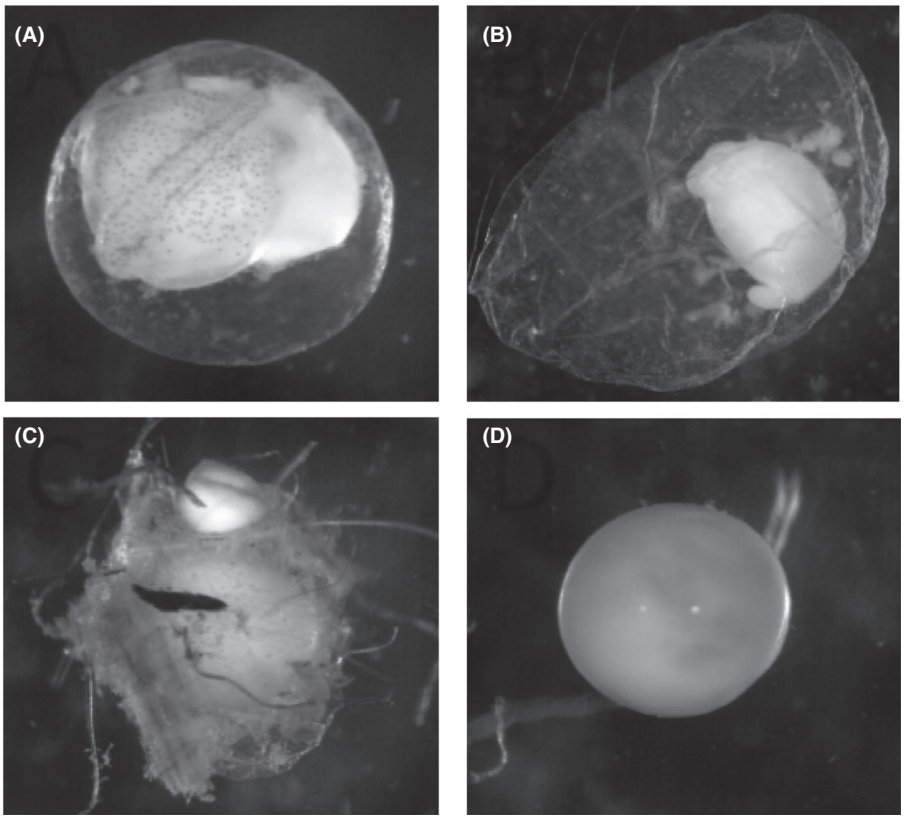


FIGURE 2. Fish eggs depicting examples of the predictor variables that were used in the random forests model. The four panels show **(A)** an egg with pigment on a definable embryo, **(B)** an egg with a deflated outer membrane and a definable embryo, **(C)** an egg with debris sticking to the outer membrane, and **(D)** an egg with no discernable embryo.

−20°C. Polymerase chain reaction was used to amplify the portions of the mitochondrial genome corresponding to the cytochrome b gene using the primers that were developed by Song et al. (1998) or cytochrome oxidase subunit I (COI) using the primers that were developed by Ivanova et al. (2007). The successfully amplified polymerase chain reaction products were sequenced and manually edited in Geneious (http://www.geneious.com; Kearse et al. 2012) and compared with the DNA sequences of known invasive carp species for positive identification. Noninvasive carp sequences were identified to species by comparing them with available data bases of DNA sequences (e.g., GenBank) or with the NCBI nonredundant database and the Megablast search algorithm (Altschul et al. 1997, as implemented in Geneious v8.1.7).

A random forests machine learning algorithm (Breiman 2001) was used to predict the taxonomic class of an individual egg from an array of predictor variables. For each egg, 13 egg morphology and four environmental metrics were recorded as predictor variables and the genetic identification was recorded as the response variable. The predictive variables were chosen as an exhaustive list of potential diagnostic egg characteristics, in which the discovery of novel relationships and/or diagnostic variables could be explored. The morphological variables included the presence or absence of pigment on the embryo, membrane deflation, debris adhered to the membrane, the presence of a definable embryo, egg development stage, average, standard deviation, and coefficient of variation of membrane and embryo diameter, late-stage embryo length, and perivitelline space index. Since fishes do not all spawn at the same time of year and have different optimum water conditions for reproduction, environmental variables such as water temperature and conductivity were measured at each site during egg collection and the ordinal day and month were included. In random forests algorithms, all of the predictor variables must have a measurement for each observation. Thus, genetically identified embryos without membranes from Grass Carp ($n = 8$), Silver Carp ($n = 44$), Bighead Carp ($n = 1$), Channel Shiner *Notropis wickliffi* ($n = 1$), Emerald Shiner *Notropis atherinoides* ($n = 2$), Striped Bass *Morone saxatilis* ($n = 2$), and White Bass *Morone chrysops* ($n = 1$) were excluded from further analysis.

Random forests does not make assumptions about the normality or independence of the data; it is applicable

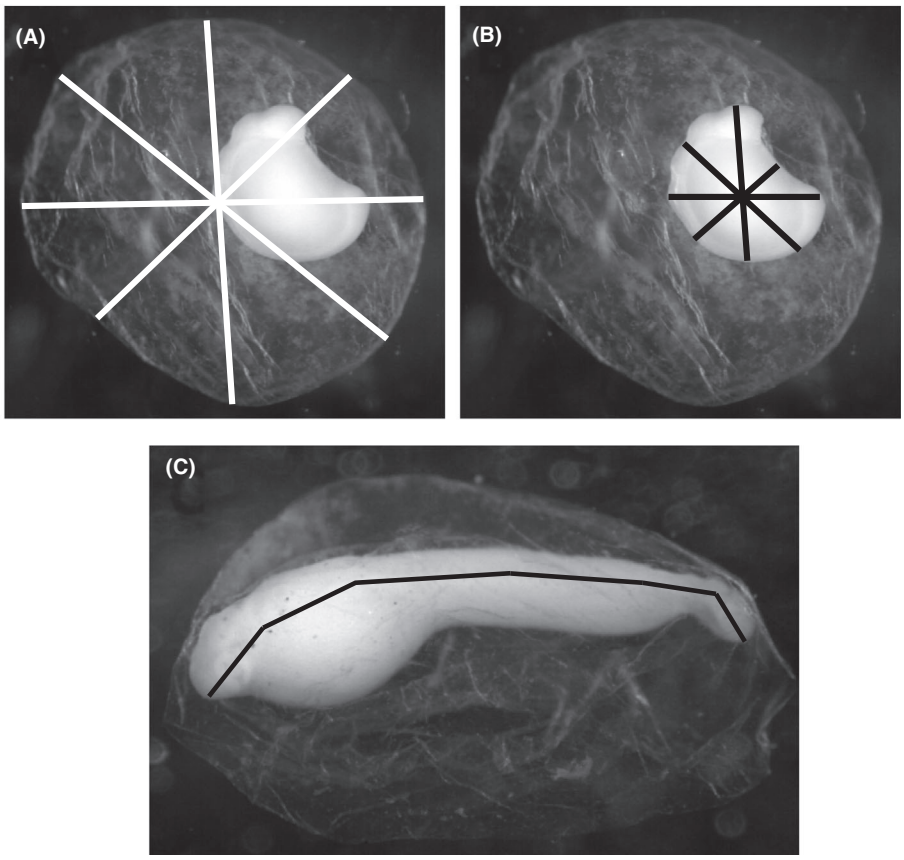


FIGURE 3. Diameter measurement placement for **(A)** outer membrane, **(B)** embryo, and **(C)** midline measurements on all of the later-stage embryos.

with continuous and categorical variables; it is relatively robust to outliers, noise, and autocorrelation; and it is a commonly used prediction model (Breiman 2001; Cutler et al. 2007; Maroco et al. 2011). Using random forests allowed for a large number of variables and many combinations of the variables to be explored for more accurately predicting egg identification at various taxonomic levels. In addition, random forests uses a random bootstrapped sample (63%) of the original data to construct 5,000 classification trees and a random subset of the predictor variables is chosen for each node (split) in a tree (Prasad et al. 2006). Combinations of three to four randomly selected predictor variables were chosen for each node. Growing a large number of trees using bootstrapped data is effective at reducing the generalization error that is associated with the model training set, and randomly selecting predictor variables at each node reduces bias and variance by decreasing the correlation of the trees. The resulting model does not overfit the data (Breiman 2001). The remaining unused original data (out-of-bag observations; 37%) for each tree are then run through the associated tree and given a predicted classification. The final prediction for each observation is the class with the most votes across all trees. Accuracies and error rates are computed using the out-of-bag predictions and averaged across all observations. There is no need to manually cross validate random forests since each tree is constructed without the out-of-bag observations, which provides the opportunity to use all of the available egg data. Random forests also internally calculates the out-of-bag error estimates that are analogous to cross validation error estimates and a classification error (Breiman 2001).

A total of five random forests models using all of the predictor variables were created to evaluate their ability to accurately identify invasive carp eggs at various taxonomic levels. Two of the models classified eggs to genus or species, and three additional models combined the eggs from Silver, Bighead, and Grass carp into a single class called invasive carp, with all of the other species grouped to family, genus, or species. For each random forest model, three metrics were calculated for each class that was associated with invasive carp from the random forests confusion matrix. The confusion matrix depicts the performance of random forests predictions compared with the identifications determined by genetics. Predictive accuracy was defined as the accuracy of the random forests models

at predicting the correct genetic identification, calculated as follows:

$$\text{Predictive accuracy} = \frac{cp_{class}}{n_{class}},$$

where, $cp_{class}$ = frequency of correct predictions of a class and $n_{class}$ = frequency of eggs genetically identified in a class. False positive error was defined as the proportion of incorrect predictions of genetic identifications from a class and was calculated as follows:

$$\text{False positive error} = \frac{WP_{class}}{N - n_{class}},$$

where $WP_{class}$ = frequency of wrong predictions of a class, $N$ = total frequency of genetic identifications from all classes. Nontarget taxa accuracy was defined as the proportion of correct predictions of nontarget classes from all nontarget genetic identifications and was calculated as follows:

$$\text{Nontarget taxa accuracy} = \frac{\sum cp_{nt.class}}{N - \sum n_{nt.class}},$$

where, $cp_{nt.class}$ = the correct predictions of a nontarget class and $n_{nt.class}$ = the frequency of genetic identifications of a nontarget class. All of the proportions were reported as percentages. All of the statistical analyses were conducted using R software (R Core Team 2013) and the "randomForest" package (Breiman 2001; Liaw and Wiener 2002). The R code and example data set for the study can be found in the Supplement available in the online version of this paper.

To reduce the number of predictor variables that were needed to accurately predict invasive carp, we used the variable importance measure to rank all of the predictor variables. A series of random forests models were created using a stepwise ascending variable introduction strategy. Random forests analysis uses an approach to measure variable importance that differs from the commonly used statistical methods for models that use Akaike's Information Criterion (Liaw and Wiener 2002). However, the use of a variable importance measure is effective at identifying predictor variables for elimination without sacrificing the model's predictive accuracy (Oh et al. 2003; Genuer et al. 2010). Random forests uses a Gini importance measure, defined as a predictor variable's degree of discriminability between classes (Oh et al. 2003). At every node of every tree, at least one of the predictor variables is used to form a split, resulting in a decrease of the splitting criterion. The Gini measure is computed as the sum of all of the decreases in the splitting criterion within the random forests trees due to a given variable, normalized by the number of trees grown (Breiman 2001). Therefore, predictor variables with low Gini measure scores are less informative for discriminating between classes and may be eliminated. The Gini measure from the species-level random forests model containing a single class for all of the invasive carp species was used to order the predictor variables based on importance. A sequence of random forests was initiated, starting with the most important variable and adding the next most important variable until all of the variables were used (Genuer et al. 2010). The class error and false positive error associated with the invasive carp class from each of the random forests was calculated and then summed to calculate the total invasive carp error. Class error was calculated as

$$\text{Class error} = 1 - \text{predictive accuracy}.$$

The random forests model with the smallest total invasive carp error was considered the most parsimonious model for efficiently predicting invasive carp egg identification.

Partial dependence plots specific to the invasive carp class were created for the final reduced predictor variable random forests. Partial dependence plots show the relative importance of each variable for predictions when the effects from all of the other variables are accounted for. Positive values have a higher influence on correctly predicting a specific class, and values near zero contribute little to accurate predictions.

## RESULTS

A total of 10,205 eggs were collected from May 5 to September 26 in 2014, and 5,929 eggs were collected from April 23 to September 25 in 2015. A subset of 2,061 eggs were measured and genetically identified. Genetic analysis successfully identified 57% (734 out of 1,294) eggs from 2014 and 71% (541 out of 767) eggs from 2015. Four species combined accounted for 83% of the identified eggs: Freshwater Drum *Aplodinotus grunniens* (32%), Silver Carp (29%), Emerald Shiner (12%), and Grass Carp (10%). The remaining 17% were composed of 25 other species, including Bighead Carp (1%). Egg membrane diameter was largest for Fathead Minnow *Pimephales promelas* (4.01 mm ± 0.71; mean ± SD) followed by Grass Carp (3.47 mm ± 0.66), Bighead Carp (3.43 mm ± 0.55), Silver Chub *Macrhybopsis storeriana* (2.97 mm ± 0.80), Silver Carp (2.84 mm ± 0.80), and Goldeye *Hiodon alosoides* (2.71 mm ± 0.55). All of the other species had average membrane diameters < 2.20 mm. Egg membrane diameter ranged from 1.79 to 4.90 mm for Grass Carp, 2.26 to 4.04 mm for Bighead Carp, and 1.46 to 4.33 mm for Silver Carp.

Genus and species random forests models with Silver, Bighead, and Grass carp evaluated in their respective taxonomic classes had low predictive accuracy and difficulty

with distinguishing invasive carp taxonomic classes from one another (Table 1). In the genus model, 56% (29 of 42) of the false negatives for *Hypophthalmichthys* were predicted as *Ctenopharyngodon* and 93% (37 of 40) of the false negatives for *Ctenopharyngodon* were predicted as *Hypophthalmichthys* (Table 1). Furthermore, 58% (37 of 64) of the false positives that were predicted as *Hypophthalmichthys* were C*tenopharyngodon* and 88% (29 of 33) of false positives that were predicted as *Ctenopharyngodon* were *Hypophthalmichthys*. In the species level model, 100% (12 of 12) of the false negatives for Bighead Carp were predicted as either Grass or Silver carp, 80% (30 of 34) of the false negatives for Grass Carp were predicted as Silver Carp, and 73% (30 of 41) false negatives for Silver Carp were predicted as either Bighead or Grass carp (Table 1). Furthermore, 100% (1 of 1) of the false positives for predicted Bighead Carp were Silver Carp, 87% (33 of 38) of predicted Grass Carp were Silver Carp and Bighead Carp, and 56% (38 of 68) of predicted Silver Carp were Bighead and Grass Carp. When Bighead, Grass, and Silver Carp were combined into a single invasive carp class and the random forests were re-run, the predictive accuracy increased and was constant across all three taxonomic levels (Table 1). Overall, random forests successfully predicted 97% of invasive carp eggs and had low invasive carp false positive rates (4–5%). The only difference between the three invasive carp random forests models was in their nontarget taxa accuracy.

The variable reduction analysis of the species random forests models with the combined invasive carp class using the decreased mean Gini scores resulted in the inclusion of 11 variables (membrane average, embryo average, deflated membrane, membrane SD, water temperature, pigment presence, ordinal day, perivitelline space index, membrane coefficient of variation, conductivity, and embryo SD) and the elimination of six predictor variables (Table 2). The reduced variable random forests incorrectly identified 2% of the invasive carp and had a 5% false positive error (Table 2). Of the 486 genetically identified invasive carp eggs that were used for the analysis, two were falsely predicted as Channel Shiner, seven were predicted as Emerald Shiner, and one was predicted as a River Shiner *Notropis blennius*. False positive predictions were distributed across nine species, but half of all 36 false positive invasive carp predictions were genetically Silver Chub *Macrhybopsis storeriana* (Table 3).

In the final random forests analysis with the reduced predictor variables and combined invasive species class, positive values for all of the partial dependence plots specific to invasive carp show that all of the variables were important for predicting invasive carp. However, their importance varies within each variable (Figure 4). For example, the importance of membrane and embryo size and variation increased with average size and variability, indicating that few other species classes had large or variable membrane or embryo size (Figure 4). However, the perivitelline space index was more important up to 0.65 and then declined, indicating there are other species classes with similar values above 0.65.

## DISCUSSION

To our knowledge, this is the first study to use a random forests machine learning algorithm to predict the identity of eggs. Visually identifying eggs is a desirable goal, but it is difficult and often error prone due to a lack of information (Richards 1985). Furthermore, morphological changes from preservation techniques render descriptions of live specimens inadequate. However, our results demonstrate that using a combination of morphometric and environmental measurements from genetically identified preserved eggs in a random forests algorithm can

TABLE 1. Invasive carp predictive accuracy and false positive rate and nontarget taxa accuracy for each taxonomic, variable, and invasive carp classification combination that was used within each random forests model.

| Taxonomic level–variables | Classes | Invasive carp | | | Nontarget taxa accuracy (%) |
|---|---|---|---|---|---|
| | | Classes | Predictive accuracy (%) | False positive rate (%) | |
| Genus–all | 17 | *Ctenopharyngodon* | 68 | 3 | 72 |
| | | *Hypophthalmichthys* | 88 | 7 | 83 |
| Species–all | 29 | Bighead Carp | 8 | 0 | 50 |
| | | Grass Carp | 73 | 3 | 71 |
| | | Silver Carp | 88 | 7 | 82 |
| Family–all | 8 | Invasive carp | 97 | 4 | 94 |
| Genus–all | 16 | Invasive carp | 97 | 4 | 93 |
| Species–all | 27 | Invasive carp | 97 | 5 | 93 |
| Species–reduced | 27 | Invasive carp | 98 | 5 | 93 |

TABLE 2. Invasive carp class error, false positive error, and total error results from the variable reduction analysis. The predictor variables were added to each subsequent model in a step-forward process based on mean decrease in Gini scores from the species random forests algorithms with Silver, Bighead, and Grass carp combined into a single class. The reduced model contained all of the variables that are marked with an asterisk.

| | Invasive carp | | |
|---|---|---|---|
| Variable added | Class error (%) | False positive error (%) | Total error (%) |
| Membrane average* | 24.49 | 14.96 | 39.44 |
| Embryo average* | 9.05 | 9.25 | 18.31 |
| Deflated membrane* | 4.32 | 9.13 | 13.45 |
| Membrane standard deviation* | 5.35 | 8.11 | 13.46 |
| Water temperature* | 4.12 | 6.59 | 10.71 |
| Pigment presence* | 3.50 | 7.10 | 10.60 |
| Ordinal day* | 2.67 | 5.07 | 7.74 |
| Perivitelline space index* | 3.09 | 5.32 | 8.41 |
| Membrane coefficient of variation* | 3.29 | 4.69 | 7.98 |
| Conductivity* | 2.47 | 4.31 | 6.78 |
| Embryo standard deviation* | 2.06 | 4.56 | 6.62 |
| Embryo coefficient of variation | 2.47 | 4.69 | 7.16 |
| Debris adhesion to membrane | 2.26 | 5.07 | 7.33 |
| Egg stage | 2.26 | 5.20 | 7.46 |
| Month | 2.26 | 5.83 | 8.09 |
| Embryo midline length | 2.88 | 4.94 | 7.82 |
| Definable embryo | 2.47 | 4.82 | 7.29 |

accurately identify eggs from the upper Mississippi River basin. Specifically, we were able to use random forests to identify invasive carp (a combined sample of Bighead, Grass and Silver carp). Random forests applications in ecological studies are limited (but see Dub et al. 2013; George et al. 2018), but its performance as a classification tool successfully met the objectives of this study.

We used a variable importance measure within random forests to determine that a multitude of morphometric and environmental characteristics were needed to differentiate invasive carp eggs from those of other upper Mississippi River fishes. Egg membrane diameter was the most important variable for identifying invasive carp eggs. However, membrane diameter alone is not diagnostic of invasive carp (George and Chapman 2013; Larson et al. 2016). Invasive carp egg membranes ranged from 1.0 to 5.5 mm and overlapped with all of the other species except Gizzard Shad *Dorosoma cepedianum*. However, the most common invasive carp membrane sizes overlapped with only a few species, such as Silver Chub and Fathead Minnow, resulting in false positive predictions. Fathead Minnow eggs were largest on average in the collection, with some approaching 5 mm, and substantially larger than the 1.4 to 1.6 mm size from live specimens, initially reported by Wynne-Edwards (1932). Egg membrane size descriptions for Silver Chub are scarce, making comparisons with other studies difficult. Goldeye also shared similar membrane sizes to invasive carp, but they could be correctly distinguished from invasive carp by using water temperature. Due to the overlap in egg membrane sizes among fishes, the addition of other egg and environmental characteristics are critical for successful identification.

The egg characteristics of Grass Carp, Bighead Carp, and Silver Carp were similar and could not be distinguished from each other in this study. Each species has a very large egg membrane diameter (Yi et al. 2006; George and Chapman 2013, 2015), but the eggs that were collected in this study had smaller membrane diameters than did those from live specimens from the Yangtze River, China (Yi et al. 2006), lower Missouri River, USA (George and Chapman 2013), and Silver Carp eggs that were collected from the Wabash River, Indiana (Lenaerts et al. 2015). Furthermore, invasive carp eggs from the upper Mississippi River that were preserved with formalin showed less variability in membrane and embryo diameter than they did in this study (Larson et al. 2016). These differences may be attributed to the desiccating properties of ethanol (Kelso and Rutherford 1996). Additionally, the variation may be attributed to the compounding factors of maternal effects (i.e., larger females produce larger eggs; George and Chapman 2013), water temperature (i.e., warmer water temperature yields larger eggs; George and Chapman 2013, 2015), and water chemistry (i.e., eggs absorb more water and become larger in soft water; Rach et al. 2010). These differences may also contribute to variation in sizes of ethanol-preserved eggs since eggs were collected from multiple watersheds and throughout each year where biotic and abiotic factors may be different. Even though average sizes may be different, the wide range of egg-size variation in all three of the invasive carp species makes differentiating among the species difficult and ineffective. Although there are some morphological differences among invasive carp during a few developmental stages (Yi et al. 2006), these do not persist through all stages. When identifying eggs based solely on morphology, a conservative approach that combines all invasive carp species into a single class should be used and subsequent genetic testing should be used to identify invasive carp eggs to species.

By not specifying a diagnostic range of values for a variable, random forests was able to create its own set of rules.

TABLE 3. Confusion matrix of genetic and random forests predicted identifications. Silver, Bighead, and Grass carp were combined into a single invasive carp class.

| Genetic identification and class | Species | Total genotyped | Predicted identification | |
|---|---|---|---|---|
| | | | Invasive carp | Nontarget species (summed) |
| | **Target species** | | | |
| Invasive carp | *Ctenopharyngodon idella*<br>*Hypophthalmichthys molitrix*<br>*Hypophthalmichthys nobilis* | 486 | 476 | 10[a] |
| | **Nontarget species** | | | |
| Banded Darter | *Etheostoma zonale* | 1 | 1 | 0 |
| Bigmouth Buffalo | *Ictiobus cyprinellus* | 7 | 1 | 6 |
| Black Buffalo | *Ictiobus niger* | 1 | 0 | 1 |
| Buffalo spp. | *Ictiobus* spp. | 10 | 0 | 10 |
| Carpsuckers spp. | *Carpiodes* spp. | 1 | 0 | 1 |
| Channel Shiner | *Notropis wickliffi* | 32 | 0 | 32 |
| Common Logperch | *Percina caprodes* | 1 | 0 | 1 |
| Common Shiner | *Luxilus cornutus* | 1 | 0 | 1 |
| Emerald Shiner | *Notropis atherinoides* | 157 | 3 | 154 |
| Fathead Minnow | *Pimephales promelas* | 5 | 4 | 1 |
| Freshwater Drum | *Aplodinotus grunniens* | 429 | 3 | 426 |
| Gizzard Shad | *Dorosoma cepedianum* | 2 | 0 | 2 |
| Goldeye | *Hiodon alosoides* | 6 | 1 | 5 |
| Quillback | *Carpiodes cyprinus* | 1 | 0 | 1 |
| River Carpsucker | *Carpiodes carpio* | 8 | 0 | 8 |
| River Shiner | *Notropis blennius* | 13 | 0 | 13 |
| Sand Shiner | *Notropis stramineus* | 1 | 0 | 1 |
| Shiner spp. | *Notropis* spp. | 33 | 0 | 33 |
| Silver Chub | *Macrhybopsis storeriana* | 36 | 18 | 18 |
| Skipjack Shad | *Alosa chrysochloris* | 1 | 0 | 1 |
| Smallmouth Buffalo | *Ictiobus bubalus* | 2 | 0 | 2 |
| Speckled Chub | *Macrhybopsis aestivalis* | 15 | 3 | 12 |
| Spotfin Shiner | *Cyprinella spiloptera* | 6 | 0 | 6 |
| Temperate Bass | *Morone* spp. | 17 | 2 | 15 |
| Walleye | *Sander vitreus* | 2 | 0 | 2 |
| White Bass | *Morone chrysops* | 1 | 0 | 1 |

[a]Genetically identified invasive carp eggs were predicted as Channel Shiner ($n = 2$), River Shiner ($n = 1$), and Emerald Shiner ($n = 7$).

Larson et al. (2016) described a single set of values for egg membrane size as a diagnostic characteristic. If the same criterion had been applied to our collections, 30% of the genetically identified invasive carp eggs would have been misclassified with a 6% false positive rate. This is a 28% higher misclassification rate and 1% higher false positive rate than was obtained with the model produced herein. For monitoring the presence of invasive carp reproduction, false positives are far less troublesome than are failures to identify a true invasive carp egg. This is especially true in areas of new reproduction when quick management responses may be necessary to curb population expansion and possible establishment. Thus, although random forests also had misclassifications and false positives, the rates were lower than those that have been obtained with other methods that have been published, providing the best method currently available for identifying eggs.

Using random forests to identify invasive carp eggs that were collected from the upper Mississippi River basin proved to be successful even though some species had few samples. Having more samples would have provided random forests with a better grasp of the variability within a given predictor variable for each species; thus, giving it greater predictive ability. Furthermore, it is beneficial to

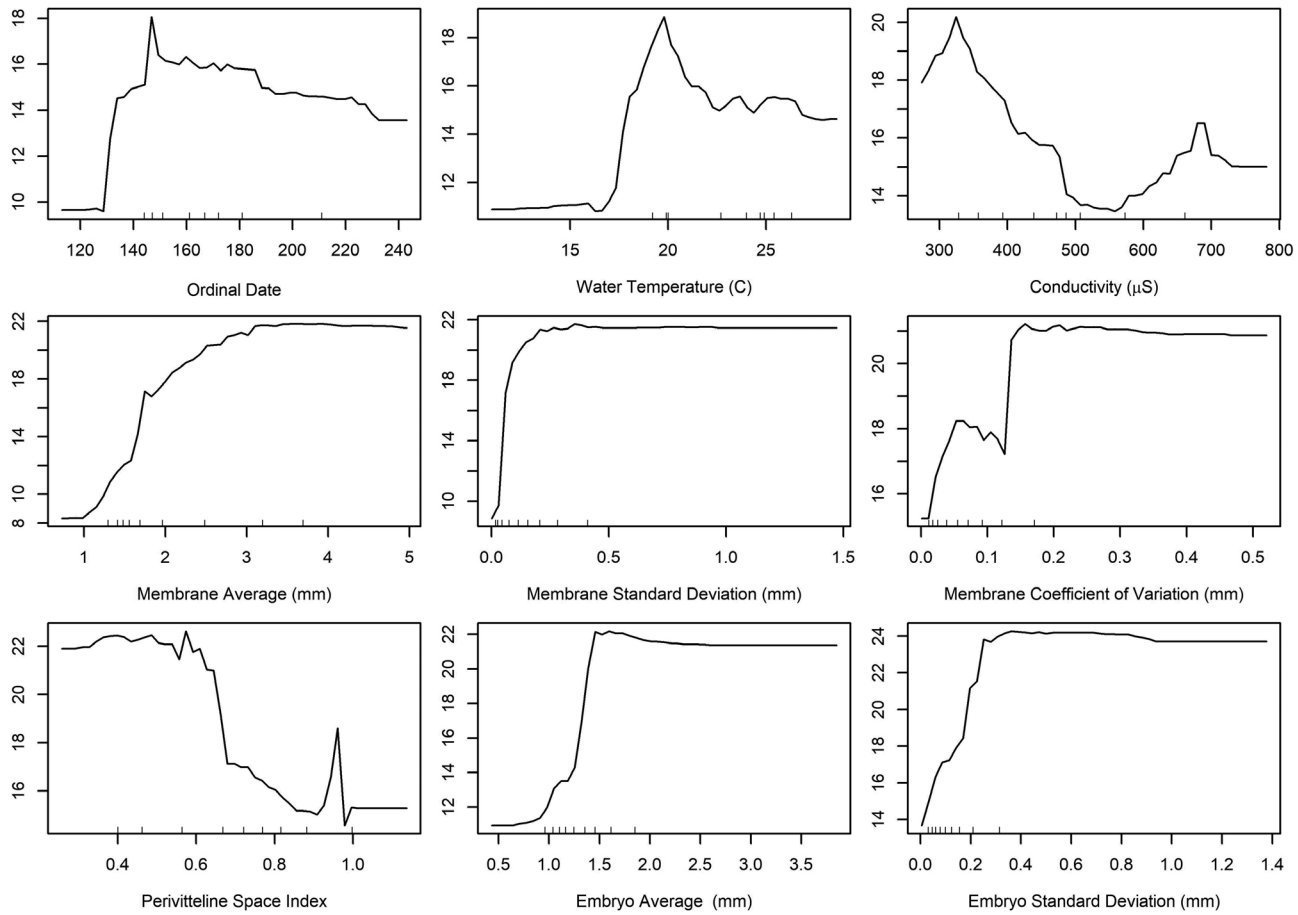


FIGURE 4. Partial dependence plots of the predictor variables from the random forests predictions that were used to identify invasive carp eggs. Partial dependence depicts the relative importance of a single variable to predicting an egg as being an invasive carp after averaging out the effects of all of the other variables. Positive values have a higher influence on correctly predicting invasive carp, and values near zero contribute minimally to accurate predictions.

have more samples in classes that are known to have overlapping values or high variation for a predictor variable (Rodriguez-Galiano et al. 2012; Brown et al. 2014). This could be achieved by including collections over many years or across geographic locations to increase sample sizes of rare or uncommon taxa. This would also increase its spatiotemporal application for identification. Random forests procedures could also be used as a prescreening tool to increase the probability that an egg that is submitted for genotyping is an invasive carp or as a backup method for identification when genetic techniques fail (e.g., only 65% of eggs that are submitted were successfully genotyped in this study). This method could be used in combination with eDNA sampling of ichthyoplankton trawls to determine which eggs to submit for genetic validation from eDNA positive trawls (Fritts et al. 2018). Regardless of application, we would suggest that any eggs that random forests has classified or misclassified as invasive carp (e.g., Silver Chub) be genetically verified to validate the results in areas where invasive carp have not yet been documented to reproduce.

It is important to note that the specifics of the random forests algorithms that were used in this study may not be directly applicable to other areas with different fish assemblages or for studies with a different objective. If a species is not in the data set, random forests will not have sufficient information to correctly identify the missing species. Future model validation with an independent data set would help to determine how broadly applicable our model results are at identifying fish eggs during different periods and from different locations. Aggregating similar groups into a single class can be useful, such as was used in this study for our target species, but may not be suitable for all studies, depending on the study objectives. Regardless, the model approach that we have outlined here allows biologists to customize random forests with abiotic and biotic features that would be effective for identifying fish eggs of other species and in other systems.

## ACKNOWLEDGMENTS

This project was funded by the Iowa Department of Natural Resources, U.S. Fish and Wildlife Service, and Iowa State University. This study was performed under the auspices of Iowa State University Institutional Animal Care and Use Committee Protocol 7-13-7599-I, and the animals were collected under state permit SC1037. The use of trade, product, or firm names is descriptive and does not imply endorsement by the U.S. Government. The authors would like to thank Nehemias Ulloa for his statistical guidance, Jim Lamer for his hospitality, Kevin Roe and Keith Turnquist for performing the genetic analyses, and all of the technicians who made this work possible. We also thank Tim Copeland and two anonymous reviewers for their helpful comments on the manuscript. There is no conflict of interest declared in this article.

## ORCID

Carlos A. Camacho,? https://orcid.org/0000-0001-9828-1987
Christopher J. Sullivan,? https://orcid.org/0000-0001-7214-3789
Michael J. Weber https://orcid.org/0000-0003-0430-3087
Clay L. Pierce https://orcid.org/0000-0001-5088-5431

## REFERENCES

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389–3402.

Becker, R. A., N. G. Sales, G. M. Santos, G. B. Santos, and D. C. Carvalho. 2015. DNA barcoding and morphological identification of neotropical ichthyoplankton from the upper Paraná and São Francisco. Journal of Fish Biology 87:159–168.

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Brown, S. C., R. E. Lester, V. L. Versace, J. Fawcett, and L. Laurenson. 2014. Hydrologic landscape regionalisation using deductive classification and random forests. PLoS (Public Library of Science) ONE [online serial] 9:e112856.

Chapman, D. C., and A. E. George. 2011. Developmental rate and behavior of early life stages of Bighead Carp and Silver Carp. U.S. Geological Survey Scientific Investigations Report 2011-5076.

Coulter, A. A., D. Keller, E. J. Bailey, and R. Goforth. 2016. Predictors of bigheaded carp drifting egg density and spawning activity in an invaded, free-flowing river. Journal of Great Lakes Research 42:83–89.

Crean, A. J., and D. J. Marshall. 2009. Coping with environmental uncertainty: dynamic bet hedging as a maternal effect. Philosophical Transactions of the Royal Society of London Series B Biological Sciences 364:1087–1096.

Cutler, D. R., T. C. Edwards Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

Deters, J. E., D. C. Chapman, and B. McElroy. 2013. Location and timing of Asian carp spawning in the lower Missouri River. Environmental Biology of Fishes 96:617–629.

Diaz-Viloria, N., L. Sanchez-Velasco, and R. Perez-Enriquez. 2005. Inhibition of DNA amplification in marine fish larvae preserved in formalin. Journal of Plankton Research 27:787–792.

Dub, J. D., R. A. Redman, D. H. Wahl, and S. J. Czesny. 2013. Utilizing random forest analysis with otolith mass and total fish length to obtain rapid and objective estimates of fish age. Canadian Journal of Fisheries and Aquatic Sciences 70:1396–1401.

Embke, H. S., P. M. Kocovsky, C. A. Richter, J. J. Pritt, C. M. Mayer, and S. S. Qian. 2016. First direct confirmation of Grass Carp spawning in a Great Lakes tributary. Journal of Great Lakes Research 42:899–903.

Freeze, M., and S. Henderson. 1982. Distribution and status of the Bighead Carp and Silver Carp in Arkansas. North American Journal of Fisheries Management 2:197–200.

Frimpong, E. A., and M. L. Henebry. 2012. Short-term effects of formalin and ethanol fixation and preservation techniques on weight and size of fish eggs. Transactions of the American Fisheries Society 141:1472–1479.

Fritts, A. K., B. C. Knights, J. Amberg, J. H. Larson, J. J. Amberg, C. Merkes, T. Tajioui, S. E. Butler, M. J. Diana, D. H. Wahl, M. J. Weber, and J. D. Waters. 2018. Development of a quantitative PCR method for screening ichthyoplankton samples for bigheaded carps. Biological Invasions 21:1143–1153.

Genuer, R., J. M. Poggi, and C. Tuleau-Malot. 2010. Variable selection using random forests. Pattern Recognition Letters 31:2225–2236.

George, A. E., and D. C. Chapman. 2013. Aspects of embryonic and larval development in Bighead Carp (*Hypophthalmichthys nobilis*) and Silver Carp (*Hypophthalmichthys molitrix*). PLoS (Public Library of Science) ONE [online journal] 8:e73829.

George, A. E., and D. C. Chapman. 2015. Embryonic and larval development and early behavior in Grass Carp, *Ctenopharyngodon idella*: implications for recruitment in rivers. PLoS (Public Library of Science) ONE [online journal] 10:e0119023.

George, E. M., M. P. Hare, D. L. Crabtree, B. F. Lantry, and L. G. Rudstam. 2018. Comparison of genetic and visual identification of Cisco and Lake Whitefish larvae from Chaumont Bay, Lake Ontario. Canadian Journal of Fisheries and Aquatic Sciences 75:1329–1336.

Hutchings, J. 1991. Fitness consequences of variation in egg size and food abundance in Brook Trout *Salvelinus fontinalis*. Evolution 45:1162–1168.

Ivanova, N. V., T. S. Zemlak, R. H. Hanner, and P. D. N. Hebert. 2007. Universal primer cocktails for fish DNA barcoding. Molecular Ecology Notes 7:544–548.

Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Kelso, W. E., and D. A. Rutherford. 1996. Collection, preservation, and identification of fish eggs and larvae. Pages 255–302 *in* B. R. Murphy and D. W. Willis, editors. Fisheries techniques. American Fisheries Society, Bethesda, Maryland.

Kolar, C. S., D. C. Chapman, W. R. Courtenay Jr., C. M. Housel, D. P. Jennings, and J. D. Williams. 2007. Bigheaded carps: a biological synopsis and environmental risk assessment. American Fisheries Society, Special Publication 33, Bethesda, Maryland.

König, U., and J. Borcherding. 2012. Preserving young-of-the-year *Perca fluviatilis* in ethanol, formalin, or in a frozen state and the consequences for measuring morphometrics. Journal of Applied Ichthyology 28:740–744.

Larson, J. H., S. G. McCalla, D. C. Chapman, C. Rees, B. C. Knights, J. M. Vallazza, A. E. George, W. B. Richardson, and J. Amberg. 2016. Genetic analysis shows that morphology alone cannot

distinguish Asian carp eggs from those of other cyprinid species. North American Journal of Fisheries Management 36:1053–1058.

Lenaerts, A., A. Coulter, Z. Feiner, and R. Goforth. 2015. Egg size variability in an establishing population of invasive Silver Carp *Hypophthalmichthys molitrix* (Valenciennes, 1844). Aquatic Invasions 10:449–461.

Liaw, A., and M. Wiener. 2002. Classification and regression by random forest. R News 2:18–22.

Lodge, D. M. 1993. Biological invasions: lessons for ecology. Trends in Ecology and Evolution 8:133–137.

Mack, R. N., D. Simberloff, W. Mark Lonsdale, H. Evans, M. Clout, and F. A. Bazzaz. 2000. Biotic invasions: causes, epidemiology, global consequences, and control. Ecological Applications 10:689–710.

Maroco, J., D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. de Mendonça. 2011. Data mining methods in the prediction of dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Research Notes [online series] 4:article 299.

Martinez, P. A., W. M. Berbel-Filho, and U. P. Jacobina. 2012. Is formalin fixation and ethanol preservation able to influence in geometric morphometric analysis? Fishes as a case study. Zoomorphology 132:87–93.

Oh, J., M. Laubach, and A. Luczak. 2003. Estimating neuronal variable importance with random forest. Institute of Electrical and Electronics Engineers, Proceedings of the 29th Annual IEEE Bioengineering Conference, New York.

Peterson, A. T., and D. A. Vieglais. 2001. Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. A new approach to ecological niche modeling, based on new tools drawn from biodiversity informatics, is applied to the challenge of predicting potential species' invasions. BioScience 51:363–371.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199.

R Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Rach, J. J., G. G. Sass, J. A. Luoma, and M. P. Gaikowski. 2010. Effects of water hardness on size and hatching success of Silver Carp eggs. North American Journal of Fisheries Management 30:230–237.

Rahel, F. J. 2002. Homogenization of freshwater faunas. Annual Review of Ecology and Systematics 33:291–315.

Richards, W. J. 1985. Status of the identification of the early life stages of fishes. Bulletin of Marine Science 37:756–760.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. 2012. An assessment of the effectiveness of a random forests classifier for landcover classification. ISPRS Journal of Photogrammetry and Remote Sensing 67:93–104.

Song, C. B., T. J. Near, and L. M. Page. 1998. Phylogenetic relations among percid fishes as inferred from mitochondrial cytochrome b DNA sequence data. Molecular Phylogenetics and Evolution 10:343–353.

Wiegand, P., J. Domhöver, and B. Brinkmann. 1996. DNA degradation in formalin fixed tissues. Pathologe 17:451–454.

Wittmann, M. E., C. L. Jerde, J. G. Howeth, S. P. Maher, A. M. Deines, J. A. Jenkins, G. W. Whitledge, S. R. Burbank, W. L. Chadderton, A. R. Mahon, J. T. Tyson, C. A. Gantz, R. P. Keller, J. M. Drake, and D. M. Lodge. 2014. Grass Carp in the Great Lakes region: establishment potential, expert perceptions, and re-evaluation of experimental evidence of ecological impact. Canadian Journal of Fisheries and Aquatic Sciences 71:992–999.

Wynne-Edwards, V. C. 1932. The breeding habits of the Black-headed Minnow (*Pimephales promelas* Raf.). Transactions of the American Fisheries Society 62:382–383.

Yi, B., Z. Liang, Z. Yu, R. Lin, and M. He. 2006. A study of the early development of Grass Carp, Black Carp, Silver Carp and Bighead Carp of the Yangtze River. Pages 15–51 *in* D. C. Chapman, editor. Early development of four cyprinids native to the Yangtze River. U.S. Geological Survey, Data Series 239, Reston, Virginia.

## SUPPORTING INFORMATION

Additional supplemental material may be found online in the Supporting Information section at the end of the article.