Evaluation of a Random Forest Model to Identify Invasive Carp Eggs Based on Morphometric Features

Katherine Goode  
*Iowa State University*, kgoode@iastate.edu

Michael J. Weber  
*Iowa State University*, mjw@iastate.edu

Aaron Matthews  
*Iowa State University*

Clay L. Pierce  
*Iowa State University*, cpierce@iastate.edu

Follow this and additional works at: [https://lib.dr.iastate.edu/nrem_pubs](https://lib.dr.iastate.edu/nrem_pubs)

Part of the Aquaculture and Fisheries Commons, Natural Resources Management and Policy Commons, Statistical Methodology Commons, and the Statistical Models Commons

The complete bibliographic information for this item can be found at [https://lib.dr.iastate.edu/nrem_pubs/374](https://lib.dr.iastate.edu/nrem_pubs/374). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).

This Article is brought to you for free and open access by the Natural Resource Ecology and Management at Iowa State University Digital Repository. It has been accepted for inclusion in Natural Resource Ecology and Management Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Evaluation of a Random Forest Model to Identify Invasive Carp Eggs Based on Morphometric Features

Abstract

Three species of invasive carp—Grass Carp *Ctenopharyngodon idella*, Silver Carp *Hypophthalmichthys molitrix*, and Bighead Carp *H. nobilis*—are rapidly spreading throughout North America. Monitoring their reproduction can help to determine establishment in new areas but is difficult due to challenges associated with identifying fish eggs. Recently, random forest models provided accurate identification of eggs based on morphological traits, but the models have not been validated using independent data. Our objective was to evaluate the predictive performance of egg identification models developed by Camacho et al. (2019) for classifying invasive carp eggs by using an independent data set. When invasive carp were grouped as one category, predictive accuracy was high at the following levels: family (89%), genus (90%), species (91%), and species with reduced predictor variables (94%). Invasive carp predictive accuracy decreased when we only considered observations from newly sampled locations (family: 9%; genus: 22%; species: 30%; species with reduced predictor variables: 70%), suggesting potential differences in egg characteristics among locations. Random forest models using a combination of previous and new data resulted in high predictive accuracy for invasive carp (96–98%) when invasive carp were grouped as one class for all models at the family, genus, and species levels. The two most influential predictor variables were average membrane diameter and average embryo diameter; the probability of predicting an invasive carp egg increased with these metrics. High predictive accuracy metrics suggest that these trained and validated random forest models can be used to identify invasive carp eggs based on morphometric variables. However, decreased performance at new locations suggests that more research would be beneficial to determine the models' applicability to a larger spatial region.

Disciplines

Aquaculture and Fisheries | Natural Resources Management and Policy | Statistical Methodology | Statistical Models

Comments

This article is published as Goode, Katherine, Michael J. Weber, Aaron Matthews, and Clay L. Pierce. "Evaluation of a Random Forest Model to Identify Invasive Carp Eggs Based on Morphometric Features." *North American Journal of Fisheries Management* (2021). doi:10.1002/nafm.10616.

Creative Commons License

This work is licensed under a Creative Commons Attribution 4.0 International License.
Evaluation of a Random Forest Model to Identify Invasive Carp Eggs Based on Morphometric Features

Katherine Goode
Department of Statistics, Iowa State University, Ames, Iowa 50011, USA

Michael J. Weber,* Aaron Matthews, and Clay L. Pierce
Natural Resource Ecology and Management, Iowa State University, Ames, Iowa 50011, USA

Abstract
Three species of invasive carp—Grass Carp *Ctenopharyngodon idella*, Silver Carp *Hypophthalmichthys molitrix*, and Bighead Carp *H. nobilis*—are rapidly spreading throughout North America. Monitoring their reproduction can help to determine establishment in new areas but is difficult due to challenges associated with identifying fish eggs. Recently, random forest models provided accurate identification of eggs based on morphological traits, but the models have not been validated using independent data. Our objective was to evaluate the predictive performance of egg identification models developed by Camacho et al. (2019) for classifying invasive carp eggs by using an independent data set. When invasive carp were grouped as one category, predictive accuracy was high at the following levels: family (89%), genus (90%), species (91%), and species with reduced predictor variables (94%). Invasive carp predictive accuracy decreased when we only considered observations from newly sampled locations (family: 9%; genus: 22%; species: 30%; species with reduced predictor variables: 70%), suggesting potential differences in egg characteristics among locations. Random forest models using a combination of previous and new data resulted in high predictive accuracy for invasive carp (96–98%) when invasive carp were grouped as one class for all models at the family, genus, and species levels. The two most influential predictor variables were average membrane diameter and average embryo diameter; the probability of predicting an invasive carp egg increased with these metrics. High predictive accuracy metrics suggest that these trained and validated random forest models can be used to identify invasive carp eggs based on morphometric variables. However, decreased performance at new locations suggests that more research would be beneficial to determine the models’ applicability to a larger spatial region.

Invasive species are increasing in abundance and spatial distribution throughout the United States via intentional or unintentional pathways (Lodge 1993; Rahel 2002). Some of the most widely recognized invasive species in North America include Grass Carp *Ctenopharyngodon idella*, Silver Carp *Hypophthalmichthys molitrix*, and Bighead Carp *H. nobilis*, collectively referred to as invasive carp. These nonnative species have rapidly expanded within the Mississippi River basin (Freeze and Henderson 1982; Nico et al. 2005; Irons et al. 2009; Wittmann et al. 2014; Camacho et al. 2021, this special section). Although the range of adults continually expands, establishment in invaded areas is determined not only by the consistent presence of adults, but also by successful reproduction as documented based on the presence of early life stages (e.g., eggs; Schrank et al. 2001; DeGrandchamp et al.)

*Corresponding author: mjw@iastate.edu
Received September 18, 2020; accepted March 9, 2021

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
The collection of eggs can provide insights into environmental conditions when reproduction is occurring; egg staging can be used to back-calculate spawning locations, and the presence of eggs but not larvae or juveniles could identify potential bottlenecks between early life stages. Therefore, understanding where invasive carp reproduction is occurring is critical for monitoring, understanding factors associated with their spread, and developing management plans (MICRA 2017).

Although it is important to understand where invasive carp reproduction is occurring, identification of invasive carp eggs is difficult due to the overlap of morphometric features with those of native species (Chapman 2006; Chapman and George 2011; George and Chapman 2013, 2015; Camacho et al. 2019). Accurate identification of invasive carp eggs in invaded areas is further complicated by their morphological plasticity relative to their native range, which is due to biotic and abiotic factors (Hutchings 1991; Mack et al. 2000; Peterson and Vieglais 2001; Crean and Marshall 2009; Lenaerts et al. 2015) and results in misidentifications and false reporting of invasive carp reproduction (USGS 2014; Larson et al. 2016). Genetic analysis affords the most accurate identification of fish eggs (Becker et al. 2015; Coulter et al. 2016; Embke et al. 2016) but is costly, limiting the number of eggs that can be identified. Therefore, an alternative, cost-effective, non-genetic identification process is essential for providing accurate identification of invasive carp eggs to determine when and where reproduction is occurring.

A recent study demonstrated the efficacy of morphological identification of invasive carp eggs by using random forest machine learning models produced from genetically identified eggs (Camacho et al. 2019). Random forest models are powerful predictive models with high accuracy (Breiman 2001). The models were trained using the characteristics of individual eggs as predictor variables (Camacho et al. 2019). Two models were produced to classify the eggs from all fish sampled to genus and species, and an additional three models were produced with Silver, Bighead, and Grass carp combined as a single class (referred to as “invasive carp”), while all other species were classified to the family, genus, or species level (Camacho et al. 2019). These models performed well (especially with the carp species grouped into one category), but because their effectiveness was only determined with the training data, it is uncertain whether the models would perform as accurately with new data or in other areas of data collection. Thus, validation of these models is desirable before application.

The objective of this study was to validate the Camacho et al. (2019) set of models for identifying invasive carp eggs from morphological characteristics. First, we reproduced the models from Camacho et al. (2019) by using data collected during 2014–2015 and validated the models by applying them to data collected at new sites during 2016 and computing validation metrics. Second, we compared the original model performance on previously sampled sites and new validation sites to gain insight into how the original models would perform when applied to data in other regions. Third, we produced a new series of models using all data from the years 2014–2016 (hereafter, “augmented models”) and compared model performance metrics and variable importance from the 2014–2015 and augmented models to determine whether the models fitted to all years of data performed well. Results from our analyses provide a series of robust, tested, and refined models that natural resource managers can use to identify eggs preserved in ethanol across several taxonomic levels based on morphological characteristics.

METHODS

Field Collection

Fish eggs were collected in major southeast Iowa rivers and the upper Mississippi River (UMR), USA, during 2014–2015 by Camacho et al. (2019) and during 2016 in this study. Sampling locations used in 2014 and 2015 included sites in the Des Moines, Skunk, and Iowa rivers (Figure 1), whereas sampling in 2016 was expanded to include the Rock and Wapsipinicon River confluences and Pools 15 and 17 of the UMR (Figure 1). Sampling of fish eggs in this study was performed in the same manner as described by Camacho et al. (2019) and was conducted every 10 d from late April through September at a series of sites at tributary confluences to the UMR or within a specific tributary reach. Confluence sites were composed of a total of three transects: (1) within the tributary, (2) within the UMR approximately 1 km upstream of the tributary mouth, and (3) within the UMR approximately 1 km downstream of the tributary mouth. At each transect, one ichthyoplankton tow was conducted in each of three habitat types: backwater, channel border, and thalweg. Backwater habitats consisted of areas with little to no flow. Channel border habitats varied in flow but were faster than backwater habitats and slower than thalweg habitats. Thalweg habitats were located in the fastest flowing portions of the river, which were typically in the center but could change depending on stream morphology. Tows collected eggs with an ichthyoplankton net (0.5-m-diameter opening; 500-µm mesh) towed just below the water surface at constant low engine RPMs or at idle in backwater habitats or if flow was negligible. Concurrently, water temperature and conductivity were measured with an ExtStik II Conductivity Meter (Extech Instruments Corp., Nashua, New Hampshire) during each ichthyoplankton tow. Tow duration was typically 4 min but was
occasionally less if debris load was high. The contents of each tow were washed into a removable cod end, drained of water in sieves, and preserved with 95% ethanol.

**Laboratory Processing**

Eggs and larvae from each tow were separated from debris by at least two individuals independently until no additional eggs were found. Samples were stored in 20-mL glass scintillation vials with 95% ethanol. When eggs were collected, the number of eggs in an individual ichthyoplankton tow ranged from 1 to 8,362. Due to the high number of eggs in some samples, subsampling was used to select eggs for genetic analysis. Subsampling consisted of selecting all eggs from ichthyoplankton samples containing 2 or fewer eggs, 1 egg from samples with 3–33 eggs, and 3% of eggs from tows containing 34 or more eggs. Eggs from each individual ichthyoplankton tow were subsampled by pouring all eggs into a clear petri dish with a numbered grid attached underneath. The petri dish was jostled in all directions until eggs appeared to be equally distributed throughout the dish. A random number generator produced a number corresponding with the grid, and an egg within the grid was removed for subsampling. This process was repeated until the set number of eggs for subsampling was met. Eggs without a membrane or embryo were excluded from the analysis due to the inability to extract material for genetic analysis or observe these physical features. If the randomly selected egg lacked a membrane or embryo, it was removed and another egg was randomly selected. For the 2014–2015 data set, Camacho et al. (2019) used this approach to genetically test 1,294 eggs in 2014 and 767 eggs in 2015 that were used in development of their model. For the 2016 validation data set, we used this same approach to randomly select 839 (5.4%) of 15,479 collected eggs for genetic identification. Each egg selected for genetic identification was photographed at 2× magnification from several angles for morphometric measurements (Olympus SZX7 microscope; Image-Pro version 7.0, Media Cybernetics, Bethesda, Maryland).

**Egg Genetics**

Each egg was stored in a separate 5-mL microcentrifuge tube with 95% ethanol for genetic analysis, which was completed at the Wisconsin Cooperative Fisheries
Unit. The DNA was extracted from eggs by using the Promega Wizard Genomic DNA Purification Kit (Promega Corp., Madison, Wisconsin), and samples were stored at −20°C. Portions of the mitochondrial genome were then amplified using PCR (Song et al. 1998) on cytochrome oxidase subunit I (Song et al. 1998; Ivanova et al. 2007). Amplified PCRs were sequenced and manually edited in Geneious (http://www.geneious.com; Kearse et al. 2012). Sequences were identified using GenBank or the National Center for Biotechnology Information database with the MegaBLAST search algorithm (Altschul et al. 1997).

Model Predictor Variables
We measured the same egg morphometrics and characteristics (13 biotic, 4 abiotic) that were described and identified as useful predictor variables by Camacho et al. (2019). These predictor variables were obtained from the pictures of each egg sent for genetic analysis and were used with the model to aid in classifying species by their unique traits (Table 1). The 13 biotic predictor variables included pigment presence or absence; deflated or nondeflated membrane; debris presence or absence; compact or diffuse embryo; egg development stage; mean, SD, and CV of embryo diameter (mm); mean, SD, and CV of membrane diameter (mm); late-stage embryo midline length (mm); and ratio of embryo to membrane average diameters (Camacho et al. 2019). Abiotic variables included water temperature (°C), conductivity (μS/cm), Julian day, and month of egg collection to account for variable spawning condition requirements (Camacho et al. 2019, 2021).

When pigment was present on the embryo (Figure 2A), it was found on the embryo mass. Deflated membranes (Figure 2B) were described as having a wrinkled and nonspherical shape. Membrane deflation is a result of preservation in alcohol; it is not known whether deflation would result from preservation in formalin. Presence or absence of debris (Figure 2C) adhesion was determined by any grassy or other organic matter fixed to the membrane of the egg. Compact embryos were identified with a clear margin between the embryo and membrane, while diffuse embryos (Figure 2D) contained an opaque distribution throughout the membrane (Kelso and Rutherford 1996). Egg development was assigned to one of eight stages (Kelso and Rutherford 1996) from a series of physical attributes that appear as the egg develops until hatching. The mean, SD, and CV of egg diameter (Figure 2E) and membrane diameter (Figure 2F) were obtained by a series of four diameter measurements 45° apart. Late-stage embryo midline length (Figure 2G) was measured on all eggs at stage 8, which occurs prior to hatching.

Statistical Analyses
Original models.—We first reproduced the models from Camacho et al. (2019), which we refer to as the “original models,” to validate the predictive accuracy using the 2016 data. Random forests construct many classification trees fitted to bootstrapped samples of the original data with a randomly selected subset of predictor variables (Breiman 2001). Due to the randomness of the bootstrap sampling and variable subsets, random forests will change every time they are fitted unless a random seed is set prior to the fitting of the model. The seed ensures the use of the same pseudorandom numbers that were used to generate the samples and subsets (Sandve et al. 2013). We used the same seed (808) as Camacho et al. (2019) to accurately reproduce the original models based on code provided in the Supplemental Materials (Camacho et al. 2020).

We fitted the original models with the same response and predictor variables as Camacho et al. (2019) that included genetically identified eggs from 2014 (734 eggs) and 2015 (541 eggs). Five of the original models were fitted using the 17 predictor variables previously described. Two

| Taxonomic level (predictor variables) | Number of classes | Class              | Predictive accuracy (%)a | False positive error (%)a | Precision (%)a |
|--------------------------------------|------------------|--------------------|--------------------------|--------------------------|----------------|
| Species (all)                        | 29               | Bighead Carp       | 8 (8)                    | 0 (0)                    | 50 (50)        |
|                                      |                  | Grass Carp         | 71 (73)                  | 3 (3)                    | 71 (71)        |
|                                      |                  | Silver Carp        | 89 (88)                  | 7 (7)                    | 82 (82)        |
| Genus (all)                          | 17               | Ctenopharyngodon   | 67 (68)                  | 3 (3)                    | 70 (72)        |
|                                      |                  | Hypophthalmichthys | 88 (88)                  | 7 (7)                    | 82 (83)        |
| Species (all)                        | 27               | Invasive carp      | 98 (97)                  | 5 (5)                    | 92 (93)        |
| Genus (all)                          | 16               | Invasive carp      | 97 (97)                  | 5 (4)                    | 93 (93)        |
| Family (all)                         | 8                | Invasive carp      | 96 (97)                  | 4 (4)                    | 94 (94)        |
| Species (reduced)                    | 27               | Invasive carp      | 98 (98)                  | 5 (5)                    | 92 (93)        |

*Metric computed in this study (metric reported by Camacho et al. 2019 is shown in parentheses).
models were fitted with the egg identification response variables of species and genus. The other three models were fitted with egg identification response variables of species, genus, and family, except that Silver, Bighead, and Grass carp were grouped as one classification category (invasive carp). The sixth model was fitted with an egg identification response variable of species with invasive carp grouped as one category and a reduced set of 11 predictor variables identified by Camacho et al. (2019) using a stepwise ascending variable introduction strategy based on the random forest variable importance: pigment presence or absence, membrane deflated or nondeflated, mean embryo diameter, SD of embryo diameter, mean membrane diameter, SD of membrane diameter, CV of membrane diameter, ratio of embryo to membrane average diameters, temperature, conductivity, and Julian day.

To verify that the original models were reproduced correctly, we computed predictive accuracy, false positive error, and precision metrics on all original models for the target classes using out-of-bag predictions to compare to the results reported by Camacho et al. (2019). Predictive accuracy was defined as the proportion of correct predictions made for a specific response variable class. False positive error was defined as the proportion of instances in which a model incorrectly predicted an observation to be in a specific response variable class. Precision was defined as the proportion of predictions identified as a specific class that were correct. The metrics were computed as

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

\[
\text{False positive error} = \frac{wp_{class}}{N - n_{class}},
\]

and

\[
\text{Precision} = \frac{cp_{class}}{n_{pred\,class}}.
\]

where \(cp_{class}\) is the number of correct predictions for a class, \(n_{class}\) is the number of eggs genetically identified in a class, \(wp_{class}\) is the number of wrong predictions of a class, \(N\) is the total number of genetic identifications from all classes, and \(n_{pred\,class}\) is the number of model predictions to a class.

**Validation of original models.**—The original models were validated using genetically identified eggs collected in 2016 to compare the predictions from the six original models to the 2016 observed genetic identifications. Predictions were compared to the observed genetic identifications by using the metrics of predictive accuracy, false positive error, and precision (previously defined) to assess the effectiveness of the original models. Confusion matrices were also constructed to visually compare the predictions (columns) made by the original models on the 2016 data to the observed genetic identifications (rows).

We also assessed how the original models performed when applied to new sites by comparing model performance of the sites in the validation data that were and were not sampled in 2014 and 2015 (Figure 1, locations represented by diamonds and stars, respectively). Model performance was assessed by computing predictive accuracy, false positive error, and precision separately for observations from sites that were only contained in the validation data and sites in the validation data that were also in the training data for the original models. The metrics were only computed for invasive carp from models with invasive carp grouped as a single response variable.

**Augmented models.**—Random forest models often improve with larger training data sets (Millard and Richardson 2015). Thus, the 2016 validation data were combined with the original 2014–2015 data, and a new series of random forest models was produced to assess how these augmented models performed with additional data. Six augmented random forest models were fitted that...
mimicked the Camacho et al. (2019) original models and used the same response and predictor variable structure as the original models described previously. The performance of the augmented models was assessed by computing predictive accuracy, false positive error, and precision using out-of-bag predictions. These metrics were compared to the original models’ metrics to gauge the similarity of the models’ performance on their respective training sets. We could not use the comparison of these metrics to determine how the two models would perform on a new set of data since they were fitted to different training sets.

The variables that were important in the original and augmented models were determined using random forest Gini importance (Breiman 1984). Importance ranks were reported instead of the actual variable importance values to avoid issues with scaling for comparison. Partial dependence plots (Friedman 2001) were created for the augmented random forest model fitted with the reduced set of variables to visualize the relationship between the classification of an egg as an invasive carp and the predictor variables. The line in a partial dependence plot represents the marginal relationship between the model prediction probability and a predictor variable (averaged across other predictor variables). The pointwise SDs from individual conditional expectation curves (Goldstein et al. 2015) were computed and included on the partial dependence plots to depict the variability in the marginal relationships. A line representing the partial dependence curve from the original model fitted with reduced variables was included on each plot for comparison. Rug plots of the predictor variable training data for the augmented model were included on the x-axes to identify regions where data were sparse.

Software and output.—Statistical analyses were conducted using R statistical software (R Core Team 2020). All random forest models were fitted using the randomForest package (Liaw and Wiener 2002), with 1,000 trees and other options set to default; partial dependence was computed using the pdp package (Greenwell 2017), and visualizations were created using the R package ggplot2 (Wickham 2016). The data, code, saved versions of the random forests, information about the R package versions and computer platform used, and examples applying the random forests are provided as Supplemental Materials in the online version of this article.

RESULTS

Original Models

The performance metrics of predictive accuracy, false positive error, and precision agreed exactly when rounded to a whole percentage or were similar to those from Camacho et al. (2019; Table 1). The largest discrepancy in metrics was a difference of 2%, which occurred twice (predictive accuracy of Grass Carp at the species level and precision of Ctenopharyngodon at the genus level). Differences for all other models were less than or equal to 1% (Table 1).

Validation Data

Of the 839 eggs sent for genetic identification from 2016, 703 were successfully identified and used as the validation data. Four species comprised the majority of the eggs (638 eggs; 91%) that were genetically identified in 2016: Freshwater Drum Aplodinotus grunniens (309 eggs; 44%), Grass Carp (158 eggs; 22%), Silver Carp (127 eggs; 18%), and Emerald Shiner Notropis atherinoides (44 eggs; 6%). The remaining eggs included five species of shiner (species unknown; 37 eggs, 5%), Speckled Chub Macrhybopsis aestivalis (13 eggs; 2%), Bighead Carp (11 eggs; 2%), River Shiner N. blennius (3 eggs; 0.4%), and Goldeye Hiodon alosoides (1 egg; 0.1%).

Validation of Original Models

When the original models were applied to the 2016 data for classification, the best performance in terms of all three metrics occurred when Bighead, Grass, and Silver carp were grouped as one category, except for the false positive error for classification of Bighead Carp at the species level (Figure 3). At the species level, predictive accuracy increased from 0% (Bighead Carp), 76% (Grass Carp), and 22% (Silver Carp) to 91% when the species were grouped as a single class of invasive carp and all predictor variables were included in the model. The predictive accuracy of the model for classifying species with the reduced set of predictor variables increased slightly to the highest predictive accuracy value for invasive carp (94%). At the genus level, predictive accuracy increased from 72% (Ctenopharyngodon) and 36% (Hypophthalmichthys) to 90% when the three invasive carp species were treated as the single class of invasive carp. Similar patterns occurred with the metrics of false positive error (0–21%) and precision (45–96%; Figure 3). The one situation in which the original model performed better with invasive carp treated as separate categories was the false positive error of 0% associated with the classification of Bighead Carp at the species level. The confusion matrix associated with the validation of this model indicated that no classifications of Bighead Carp were made by the model, even for the 11 genetically identified Bighead Carp (Figure 4A), resulting in the 0% false positive error.

Confusion matrices from the application of the original models to the 2016 validation data showed that the misclassifications of Bighead, Grass, and Silver carp were due to classification as another invasive carp (Figure 4). At the species level, 100% (11 of 11) of Bighead Carp misclassifications were classified as Grass or Silver Carp, 76% (29 of 38) of Grass Carp misclassifications were classified as
Silver Carp, and 99% (98 of 99) of Silver Carp misclassifications were classified as Grass Carp (Figure 4A). At the genus level, 80% (36 of 45) of Ctenopharyngodon misclassifications were classified as Hypophthalmichthys and 99% (88 of 89) of Hypophthalmichthys misclassifications were classified as Ctenopharyngodon (Figure 4B). For classifications obtained when the target species were grouped as invasive carp, misclassifications were often classified as a taxon level within the same family. At the species level, 89% (24 of 27) of invasive carp misclassifications were classified as Silver Chub with all predictor variables (Figure 4C) and 83% (15 of 18) of invasive carp misclassifications were classified as Silver Chub with reduced predictor variables (Figure 4E). At the genus level, 93% (27 of 29) invasive carp misclassifications were classified as Macrhubopsis (Figure 4D). At the family level, 97% (33 of 34) invasive carp misclassifications were classified as Cyprinidae (Figure 4F).

For the assessment of model performance on new sites, the sites that were only sampled in 2016 had 23 invasive carp eggs and the sites that were previously sampled for the training data had 273 invasive carp eggs. The original models performed much better on observations from sampling locations also included in the training data in terms of predictive accuracy and precision and were similar in terms of false positive error (Figure 5). When all predictor variables were used, the predictive accuracy and precision were on average 75% and 41% better, respectively, for previously sampled locations across taxonomic levels. When the reduced set of variables was used, the differences in predictive accuracy and precision were reduced to 26% and 12%, respectively.

**Augmented Models**

Performance metrics of the augmented models followed a pattern similar to that of the original model metrics computed on the respective training data sets for all response variables (Figure 6). The largest discrepancy in model performance (50% difference) occurred with the precision computed at the species level for Bighead Carp, where precision was 50% for the original model and 0% for the augmented model; however, there were only two

---

**FIGURE 3.** Validation metrics of predictive accuracy, false positive error, and precision (reported as percentages) for the original models computed using the 2016 data for each of the invasive carp classification categories. The precision for Bighead Carp at the species level was not computable because the model produced no classifications of Bighead Carp.
FIGURE 4. Heat maps depicting the confusion matrices for predictions from the original random forest models on the 2016 validation data (IC = invasive carp). The rows represent the observed genetic identifications, and the columns represent the random forest model predictions. Cells on the diagonal (top left to bottom right) represent the correct model predictions. The shade of a cell represents the number of eggs.
classifications of Bighead Carp in the original model and one in the augmented model. There was not a clear pattern indicating that one set of models performed better when applied to the respective training data. Predictive accuracy was highest for both model sets when Bighead, Grass, and Silver carp were grouped as invasive carp at the species, genus, and family levels.

Variable importance ranks of the augmented and original models had similar patterns across and within the two sets of models (Figure 7). Average membrane diameter and average embryo diameter were the first and second most important variables in all models. Deflated membrane, membrane diameter SD, and pigment presence were ranked as the third through fifth most important variables in most models (the order of these three variables changed depending on the model), whereas month, larval length, and compact versus diffuse embryo were ranked as the three least important variables in all models that included them as predictors.

Partial dependence plots of the original and augmented models for classifying species with invasive carp as one category and a reduced set of predictor variables indicated similar marginal relationships between the classification probability of an invasive carp egg and the predictor variables for both models (Figure 8). Partial dependence plots indicated an increase in classification probability of an invasive carp egg with an increase in average membrane diameter, average embryo diameter, membrane diameter SD, ratio of embryo to membrane average diameters, Julian day, and temperature. For many of these variables, there was a particular range in the predictor variable where the increase occurred in the random forest probability. For instance, the average probability of an egg being classified as an invasive carp increased by 27% when mean membrane diameter increased from 1.5 to 2.5 mm (Figure 8A) and the probability increased by 29% when mean embryo diameter increased from 1 to 2 mm (Figure 8B). Additionally, eggs without the presence of pigment and eggs with deflated membranes had higher average classification probabilities as invasive carp (Figure 8C, E).

DISCUSSION

Few tools beyond genetic analysis are available to identify fish eggs, making it difficult to assess reproduction by
particular taxa. Recently, the use of random forest machine learning has been proven effective for identifying eggs, even after the eggs were morphologically distorted from ethanol preservation (Camacho et al. 2019). Our results indicate that the original models performed well when applied to the validation data for classifying invasive carp at the species, genus, and family levels. Thus, the use of egg morphometrics for identifying invasive carp eggs is a valuable tool that provides a quick and inexpensive method for assessing areas of reproduction that could subsequently warrant management actions.

For both the validation of the original models from Camacho et al. (2019) and the performance of the original and augmented models on the corresponding training data, predictive accuracy, false positive error, and precision metrics indicated good performance at the species, genus, and family levels. Thus, the use of egg morphometrics for identifying invasive carp eggs is a valuable tool that provides a quick and inexpensive method for assessing areas of reproduction that could subsequently warrant management actions.

For both the validation of the original models from Camacho et al. (2019) and the performance of the original and augmented models on the corresponding training data, predictive accuracy, false positive error, and precision metrics indicated good performance at the species, genus, and family levels. Thus, the use of egg morphometrics for identifying invasive carp eggs is a valuable tool that provides a quick and inexpensive method for assessing areas of reproduction that could subsequently warrant management actions.

Grass, and Silver carp were treated as separated classes at the species and genus levels. The two most important predictors in the random forest models were average membrane diameter and average embryo diameter. However, the size of these characteristics was very similar among the three species (average membrane diameter [mean ± SE]: 3.1 ± 0.002 mm for Silver Carp, 3.7 ± 0.030 mm for Bighead Carp, 3.6 ± 0.002 mm for Grass Carp; average embryo diameter [mean ± SE]: 1.7 ± 0.001 mm for Silver Carp, 2.0 ± 0.021 mm for Bighead Carp, 1.6 ± 0.001 mm for Grass Carp). Consequently, there was not enough variation in these two key morphological characteristics to distinguish among the three invasive carp species compared to other species that had smaller membrane and embryo diameters. In some instances, simply knowing that reproduction of this group of invasive fish is occurring within a given area may be sufficient, particularly in newly invaded areas where reproduction has not been previously documented. When species-level identification is necessary,
this random forest model still provides a useful initial screening tool to identify invasive carp eggs that can then be further identified using genetic analysis, thereby saving considerable time and resources compared to genetically identifying all eggs collected.

When the validation data were separated into two data sets based on whether an egg was collected from a site sampled prior to 2016 or not, the original models performed well at classifying the category of invasive carp for sites previously sampled but were less accurate for the new sites at all taxonomic levels. The number of invasive carp eggs collected at new sites in the validation data set was low \( n = 23 \), so it is not clear whether the decrease in performance was due to observations being from sites not included in the training data or due to small sample size. However, the original model for classifying species with the reduced set of predictor variables outperformed the other three models on the new sites, suggesting that the reduced number of predictor variables may have helped to prevent overfitting. Data from more sites and a larger geographic region would have to be collected to better assess model performance on sites not included in the training data and to evaluate widespread use of the models, but the initial results from this analysis indicate that using a smaller set of predictor variables obtained by a selection process may reduce overfitting to provide better predictive performance. We caution against the application of these models to other river systems without prior validation of the models since egg characteristics may differ geographically. Additionally, retraining of the models with data from other regions may be necessary to account for different fish species in those regions that are not represented in our training data.

Average membrane diameter, average embryo diameter, a deflated membrane, membrane diameter SD, and pigment presence or absence were consistently found to be important predictor variables across models. Large egg size has been suspected to be an important characteristic of invasive carp eggs (Yi et al. 2006; George and Chapman 2013, 2015), but sole reliance on this metric previously resulted in false positive detections that led to inaccurate reporting of new areas of invasive carp reproduction (Larson et al. 2016). Additionally, invasive carp egg characteristics can vary spatially, between native and invaded locations, and with maternal size (George and Chapman 2013, 2015; Lenaerts et al. 2015), making it
difficult to correctly identify them in newly invaded regions. Our results suggest that the combination of these predictor variables within a random forest framework can be successfully used to identify invasive carp eggs.

Although random forest analysis is a powerful approach for prediction, there are several limitations that must be considered. First, random forests, along with many machine learning models, are difficult to interpret due to the complicated algorithmic nature of the model, unlike more traditional statistical methods, such as linear models (Adadi and Berrada 2018; Gilpin et al. 2018; Guidotti et al. 2018). Variable importance and partial dependence plots provide insight into the relationships between a model prediction and the predictor variables, but both methods are known to suffer from bias when predictor variables are correlated (Strobl et al. 2007). There was moderately high correlation between some predictor variables in the reduced set (Figure 9; pairs of variables with three largest correlation magnitudes [Pearson’s product-moment correlation coefficient]: membrane diameter SD and membrane diameter CV [0.87]; average membrane diameter and ratio of embryo to membrane average diameters [−0.75]; temperature and Julian day [0.67]). Future work could be accomplished using methods that adjust for correlation among the variables to better understand the relationships between the random forest probabilities of invasive carp classifications and predictor variables. Additionally, methods exist and are being developed that could be used to explain the prediction for a single observation of interest (Adadi and Berrada 2018; Gilpin et al. 2018; Guidotti et al. 2018). These methods could be used to better understand instances when a random forest made an incorrect classification. A second limitation of random forest models is the inability to produce a simple predictive formula. Thus, egg morphological characteristics cannot be measured and easily input into a mathematical
equation to predict the probability that an egg is an invasive carp. The random forest algorithm presented here can be used to identify eggs in other studies by following an approach similar to that we have outlined, but it can be challenging to implement. In the future, it would be possible to create a more user-friendly interface, such as an R Shiny application (Chang et al. 2021), that users could access via a Web browser and input the morphometric features, and a classification (and corresponding probability) of the egg as an invasive carp egg would be returned. Regardless of these limitations, the work presented here highlights the potential for random forest models to identify invasive carp eggs quickly and inexpensively, providing a rapid assessment of reproduction.

ACKNOWLEDGMENTS

We thank C. Camacho, C. Sullivan, and the numerous undergraduate research technicians from Iowa State University who helped complete field and laboratory work. We are grateful to Amy George for comments that improved the manuscript. This study was funded by the Iowa Department of Natural Resources (Contract 14CRDFBGSCHO-0001) and the U.S. Fish and Wildlife Service (Contracts F16AP00791 and F17AP00923). Work was performed under the auspices of the Institutional Animal Care and Use Committee at Iowa State University (Protocol Permit 7-13-7599-I), and animals were collected under Iowa State Permit SC1037. There is no conflict of interest declared in this article.

ORCID

Katherine Goode https://orcid.org/0000-0003-2863-4470
Michael J. Weber https://orcid.org/0000-0003-0430-3087
Clay L. Pierce https://orcid.org/0000-0001-5088-5431

REFERENCES

Adadi, A., and M. Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160.
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. 1984. Classification and regression trees. Chapman and Hall, New York.
Breiman, L. 2001. Random forests. Machine Learning 45:5–32.
Camacho, C. A., C. J. Sullivan, M. J. Weber, and C. L. Pierce. 2019. Morphological identification of Bighead Carp, Silver Carp, and Grass Carp eggs using random forests machine learning classification. North American Journal of Fisheries Management 39:1373–1384.

Camacho, C. A., C. J. Sullivan, M. J. Weber, and C. L. Pierce. 2021. Invasive carp reproduction phenology in tributaries of the upper Mississippi River. North American Journal of Fisheries Management. DOI: 10.1002/nafm.10499.

Camacho, C. A., M. J. Weber, and C. L. Pierce. 2020. Random forest analysis to identify invasive carp from the upper Mississippi River, Iowa [online data set]. Iowa State University, Ames. Available: https://doi.org/10.25380/iastate.9879716.v1. (September 2020).

Chang, W., J. Cheng, J. J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges. 2021. shiny: web application framework for R. R package version 1.6.0. Available: https://CRAN.R-project.org/package=shiny. (April 2021).

Chapman, D. C. 2006. Early development of four cyprinids native to the Yangtze River, China. U.S. Geological Survey, Data Series 239, Reston, Virginia.

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 2011-3076, Reston, Virginia.

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 42:89–90.

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

DeGrandchamp, K. L., J. E. Garvey, and L. A. Csoboth. 2007. Linking terrestrial inputs and invasive species invasion to aquatic ecosystems. North American Journal of Fisheries Management 27:197–200.

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

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.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29:1189–1232.

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 serial] 8(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 serial] 10(3):e0119023.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining explanations: an overview of interpretability of machine learning. Pages 80–89 in F. Bonchi and F. Provost, editors. 2018 IEEE 5th international conference on data science and advanced analytics. Institute of Electrical and Electronics Engineers, New York.

Goldstein, A., A. Kapelman, J. Bleich, and E. Pitkin. 2015. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24:44–65.

Greenwell, B. 2017. Pdp: an R package for constructing partial dependence plots. The R Journal 9:421–436.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti. 2018. A survey of methods for explaining black box models. ACM (Association for Computing Machinery) Computing Surveys 51(5):article 93.

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

Irons, K. S., S. A. DeLain, E. Gittinger, B. S. Ickes, C. S. Kolar, D. Ostendorf, E. N. Ratcliff, and A. J. Benson. 2009. Nonnative fishes in the upper Mississippi River system. U.S. Geological Survey, Scientific Investigations Report 2009-5176, Reston, Virginia.

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.

Kears, M., R. Moir, A. Wilson, S. Stonen-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierier, B. Ashton, P. Meinjes, 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 R. B. Murphy and D. W. Willis, editors. Fisheries techniques, 2nd edition. American Fisheries Society, Bethesda, Maryland.

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. Ambreg. 2016. Genetic analysis shows that morphology alone cannot distinguish invasive carp egg 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 established population of invasive Silver Carp Hypophthalmichthys molitrix. Aquatic Invasions 10:449–461.

Lia, 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. M. Lonsdale, H. Evans, M. Clout, and F. A. Bazzaz. 2000. Biotic invasions: causes, epidemiology, global consequences, and control. Ecological Applications 10:689–710.

MICRA (Mississippi Interstate Cooperative Resource Association). 2017. Monitoring and response plan for Asian carp in the Mississippi River basin. MICRA, Carbondale, Illinois.

Millard, K., and M. Richardson. 2015. On the importance of training data sample selection in random forest image classification: a case study in peatland ecosystem mapping. Remote Sensing (Basel, Switzerland) 7:8489–8515.

Nico, L. G., J. D. Williams, and H. L. Jelks. 2005. Black Carp: biological synopsis and risk assessment of an introduced fish. American Fisheries Society, Special Publication 32, Bethesda, Maryland.

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.

R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available: https://www.R-project.org/. (April 2021).

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

Sandve, G. K., A. Nekrutenko, J. Taylor, and E. Hovig. 2013. Ten simple rules for reproducible computational research. PLOS (Public Library of Science) Computational Biology [online serial] 9:e1002858.
Schrank, S. J., C. S. Guy, M. R. Whiles, and B. L. Brock. 2001. Influence of instream and landscape-level factors on the distribution of Topeka Shiners Notropis topeka in Kansas streams. Copeia 2001:413–421.

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.

Strobl, C., A. L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC (BioMed Central). Bioinformatics 8:article 25.

USGS (U.S. Geological Survey). 2014. Genetic analysis reveals fish eggs found in upper Mississippi River are not invasive carp. USGS, Reston, Virginia.

Wickham, H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York.

Wittmann, M. E., C. L. Jerde, J. G. Howeth, S. P. Maher, A. M. Deines, J. A. Jenkins, G. W. Whitlege, 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.

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.