Estimating Life-Stage Specific Habitat Capacity for Anadromous Salmonids using Quantile Random Forest Models

Kevin E. See[1,*], Michael W. Ackerman[1], Richard A. Carmichael[1], Sarah L. Hoffmann[1], and Chris Beasle

February 04, 2020

Abstract

Needs more...

The QRF model presented here provided estimates of habitat carrying capacity for Chinook salmon parr during the summer months, at both the site and watershed scale. Total capacity estimates for watersheds closely matched estimates from alternative fish productivity models. Carrying capacity estimates based on QRF, like those presented here, provide managers a framework to guide the identification, prioritization, and development of habitat rehabilitation actions to recover salmon populations.

Contents

[1] Biomark, Inc. 705 South 8th St., Boise, Idaho, 83702, USA

[*] Correspondence: Kevin E. See <Kevin.See@biomark.com>

1 Introduction

The decline of anadromous Pacific salmonids (*Oncorhynchus spp.*) across the Pacific Northwest, USA has prompted numerous actions aimed at reversing that trend. These actions are often categorized into four H's – harvest modification, hatchery practices, hydro-system operations, and habitat rehabilitation. Problematically, there is substantial uncertainty regarding the degree of change that can be exerted across and within these categories, and what combination of changes will most cost-effectively and sustainably reduce mortality. Recently released "de-listing" criteria (NOAA 2016 ***need reference details***) identified adult escapement targets at the population scale, providing a quantitative metric useful for evaluating the magnitude of survival improvements required. These abundance targets provide a benchmark against which habitat rehabilitation actions can be measured. Here we describe an approach for estimating life-stage specific habitat-based carrying capacity that can be used to quantitatively identify the magnitude of tributary habitat rehabilitation necessary to support de-listing. For perhaps the first time, the necessity of tributary habitat restoration can be demonstrated and the magnitude of required change can be placed in context with the other "H's."

Pacific salmon (*Oncorhynchus spp.*) species have experienced large declines in abundance throughout much of their range (Good et al. 2005). Declines can be partially attributed to lost or altered habitat, and thus, efforts to recover depleted salmon populations are replete with efforts to rehabilitate habitat used during the freshwater life-stages. Specifically, restoring salmonid carrying capacity through tributary rehabilitation actions has been identified as a key component of recovery efforts for salmon and steelhead (*Oncorhynchus mykiss*) in the Pacific Northwest, USA. Efforts have included increasing and improving existing habitat for both spawning adults and rearing juveniles. However, estimating habitat carrying capacity, both historic and contemporary, for various life-stages of Pacific salmon, as well as identifying important habitat characteristics that influence capacity, has been an ongoing but necessary challenge. Reliable methods to better understand fish-habitat relationships as well as to estimate capacity are necessary to identify those salmon and steelhead life-stages that are limited by habitat capacity, and further, to better direct tributary rehabilitation efforts.

Fausch et al. (1988) conducted a thorough review of attempts to predict the standing crop of fish from measurable habitat covariates from 1950 to 1985, and found that the vast majority of multiple linear regression models failed to detect a significant fish-habitat signal. Since that review, there has been progress in identifying some fish-habitat relationships for some salmonid species. Nickelson et al. (1992) found that juvenile coho were found in higher densities within pool habitat on the Oregon coast. Similarly, pool and pond densities were good predictors of coho smolt densities in western Washington (Sharma and Hilborn 2001). Bryant and Woodsmith (2009) found that juvenile coho abundance was positively related to large wood at the reach scale, however their results demonstrated a negative relationship between abundance and the number of pools. Braun and Reynolds (2011) similarly found positive associations between spawner densities of sockeye in the Fraser River and large wood, in addition to positive relationships to percent undercuts and percent pools. Densities of adult spawning coho were also higher in forested areas compared to urban or agricultural areas in the Snohomish River watershed (Pess et al. 2002). Mossop and Bradford (2006) examined juvenile Chinook in the Yukon river and found positive correlations between the log of fish density and several metrics related to residual pool dimensions and large woody debris abundance as well as a negative correlation between fish density and gradient. These studies were focused on predicting observed fish densities, not necessarily capacity, and for most of them the predictive extent is limited to a particular watershed. In addition, they all assumed some form of linear fish-habitat relationship, which often results in weak predictive power.

A number of studies have utilized other modeling approaches to elicit non-salmonid fish habitat relationships. Dunham et al. (2002) used a quantile regression approach to show a negative relationship between cutthroat trout densities and the width:depth ratio of a stream for the upper quantiles of trout density. The same approach was also used to map the potential extent of sole in the English Channel and southern North Sea (Eastwood et al. 2003). Machine learning models such as boosted regression trees and random forests have been used to examine species biomass, diversity and distribution for a number of different species (Pittman

et al. 2009, Knudby et al. 2010, and Compton et al. 2012). The results from these studies highlight the importance and effectiveness of using techniques that can accommodate non-linear fish habitat relationships.

Most studies that have investigated fish habitat relationships focus on predicting a species' distribution (presence / absence) or the average abundance or density, neither of which can be easily manipulated to predict carrying capacity. Further, many of these studies focus on only one or two measures of habitat. Sweka and Mackey (2010) estimated carrying capacity of Atlantic salmon parr using a quantile regression approach, but the only habitat covariate they included was cumulative drainage area. Traditionally, carrying capacity for salmonids has been estimated through stock-recruitment curves. However, this requires a long time-series of data with variety in the number of spawners which is not usually available (Cramer and Ackerman 2009).

In fisheries it has long been recognized that that biotic and abiotic factors limit productivity within and across life-stages. For the purposes of this paper, we define carrying capacity as the maximum number of individuals that can be supported given the quantity and quality of habitat available at a given life-stage. We assume that higher observed relative densities within a given life stage are a function of habitat quantity and quality. Furthermore, we assert that observed fish density is a poor predictor of habitat capacity owing to both a paucity of individuals, especially for threatened or endangered species, and the existence of unmeasured variables that may serve to limit capacity. To address this, we have developed a model to estimate juvenile rearing capacity for Pacific salmon in wadeable streams based on quantile regression forest (QRF) (Meinshausen 2006) models using measurements of fish abundance (and density) and habitat characteristics. QRF models combine the theory and justification of quantile regression modeling (Koenker and Bassett Jr 1978, Cade and Noon 2003) with the flexibility and framework of random forest models (Breiman 2001). They account for unmeasured variables and can be used to describe the entire distribution of predicted fish densities for a given set of habitat conditions, not just the mean expected density. Random forest models have been shown to outperform more standard parametric models in predicting fish-habitat relationships in other contexts (Knudby et al. 2010). Quantile regression forests share many of the benefits of random forest models, such as the ability to capture non-linear relationships between independent and dependent variables, naturally incorporate interactions between covariates, and work with untransformed data while being robust to outliers (Prasad et al. 2006). Meanwhile, quantile regression models have been used in a variety of ecological systems to estimate the effect of limiting factors (Terrell et al. 1996, Cade and Noon 2003).

The fish abundance/density and habitat data used to fit the QRF model presented here were available from seven watersheds within the interior Columbia River Basin (CRB), Pacific Northwest, USA. Within the interior CRB two major runs of Chinook salmon (*Oncorhynchus tshawytscha*; hereafter Chinook salmon) occur, stream-type (i.e., spring/summer run) and ocean-type (i.e., fall run), each characterized by different life history characteristics. Stream-type Chinook salmon enter freshwater earlier in the year, spawn in the upper reaches of a watershed, and the juveniles rear for up to 16 months in the freshwater before entering the ocean as smolts. Ocean-type Chinook salmon enter freshwater later (e.g. fall or winter) spawn lower in the watershed, and the juveniles may only spend between several weeks and six months in freshwater before migrating to the ocean. Here we focus on juvenile stream-type Chinook, in particular the summer rearing period during low flow. Data presented here are from Chinook salmon populations in the Upper Columbia River spring-run and Snake River spring/summer-run Evolutionary Significant Units (ESU). The Upper Columbia Spring-run ESU is listed as endangered under the Endangered Species Act, the Snake River Spring/Summer-run is listed as threatened (***need citation***).

In this study, we developed a QRF model to

1. Identify measured habitat characteristics that are most strongly associated with observed Chinook salmon parr abundance and density,
2. Use paired fish and habitat measurements to elicit fish-habitat relationships for those habitat characteristics identified as important for determining fish abundance and density,
3. Predict contemporary habitat carrying capacity at all sites where the important habitat characteristics are measures (i.e., CHaMP sites within the Columbia River Basin),
4. Extrapolate capacity predictions at CHaMP sites across a watershed using globally available attribute data to estimate the Chinook salmon parr capacity of that watershed, and

5. Validate estimates of carrying capacity from QRF across multiple watersheds using independent estimates of capacity (e.g., spawner-recruit relationships).

This manuscript incorporates multiple measures of stream habitat to estimate fish-habitat relationships that encompass the collinear nature of most stream habitat metrics and can be used to predict carrying capacity. Our approach moves across several spatial scales, inferring fish-habitat relationships from detailed, localized habitat data and extrapolating capacity predictions across wide swaths of unsampled locations.

2 Methods

2.1 Study Site

Fish and habitat data used in our study were collected from eleven watersheds within the interior Columbia River Basin (CRB), Pacific Northwest, USA (Figure 1). The CRB covers more than 668,000 km^2 draining large portions of Idaho, Oregon, and Washington, and smaller portions of Montana, Nevada, Utah, and Wyoming, as well as the southeastern portion of British Columbia. The habitat data used to populate the QRF model were collected by the Columbia Habitat Monitoring Program (CHaMP) (Volk et al. 2017) and were downloaded from https://www.champmonitoring.org. Data from the following eleven CHaMP watersheds were used in this study: Asotin, Entiat, John Day, Lemhi, Methow, Minam, South Fork Salmon, Tucannon, Upper Grande Ronde, Wenatchee and Yankee Fork. Juvenile density and abundance data were collected in a subset of seven watersheds, at CHaMP survey reaches and were graciously provided by a number of projects, including the Integrated Status and Effectiveness Monitoring Project (Volk et al. 2017).

2.2 Data

CHaMP sites are 200 m to 500 m reaches within wadeable streams across select basins within the interior Columbia River Basin (CRB) selected based on a spatially balanced Generalized Random Tesselation Stratified (GRTS) sample selection algorithm (Stevens Jr and Olsen 1999, 2004). Habitat data within CHaMP sites were collected using the CHaMP protocol (CHaMP (Columbia Habitat Monitoring Program) 2016) which calls for field data collection during the low-flow period, typically from June through October. CHaMP habitat data include, but are not limited to, measurements describing channel complexity, channel units, disturbance, fish cover, large woody debris, riparian cover, size (depth, width, discharge), substrate, temperature, and water quality.

Juvenile fish surveys were conducted for sp/sum Chinook salmon parr during the summer low-flow season at many of the same sites surveyed using the CHaMP protocol. Survey methods included mark-recapture, three-pass removal sampling, two-pass removal sampling, and single-pass electrofishing, as well as snorkeling. These data were used to estimate sp/sum Chinook salmon parr abundance at all CHaMP sites where fish survey data were available. Three-pass removal estimates used the Carle-Strub estimator (Carle and Strub 1978), following advice from Hedger et al. (2013). Two-pass removal estimates used the estimator described by Seber (2002). Mark-recapture estimates used Chapman's modified Lincoln-Peterson estimator (Chapman 1951) and were deemed valid if they met the criteria described in Robson and Regier (1964). These estimates were made using the *removal* function from the *FSA* package (Ogle et al. 2019) or the *closedp.bc* function from the *Rcapture* package (Rivest and Baillargeon 2019) in R software (R Core Team 2019). Snorkel counts were transformed to abundance estimates using paired snorkel-electrofishing sites to calibrate snorkel counts. For sites with invalid estimates or that were sampled with a single electrofishing pass, we developed an estimate of capture probability based on valid estimates, using a binomial generalized linear mixed effects model. Fixed effects were species, wetted width of the site, density of fish caught on the first pass and all possible two-way interactions. We included a random effect for fish crew / watershed. We used this model to predict abundances based on the number of fish caught on the first pass and any other covariates.

Abundance estimates at all sites were then translated into linear (parr/m) fish densities which were paired with the associated CHaMP habitat data. For sites that were sampled in multiple years, only the fish and habitat data from the year with the highest observed fish density was retained to avoid possible pseudo-replication, while remaining consistent with our goal of estimating carrying capacity. After removing duplicate samples,

our initial dataset contained 328 unique sites with paired fish-habitat data (Table **??**).

2.3 Habitat Covariate Selection

A key step in developing a QRF model to predict fish capacities was selecting the habitat covariates to include in the model. Random forest models naturally incorporate interactions between correlated covariates, which is essential since nearly all habitat variables are considered correlated to one degree or another, however, we aimed to avoid overly redundant variables (i.e., variables that measure similar aspects of the habitat). Further, including too many covariates can result in overfitting of the model (e.g., including as many covariates as data points).

To prevent overfitting, we pared down the more than 100 metrics generated by the CHaMP protocol describing the quantity and quality of fish habitat for each survey site. To assist in determining the habitat metrics to include in the QRF model, we used the Maximal Information-Based Nonparametric Exploration (MINE) class of statistics (Reshef et al. 2011) to determine those habitat characteristics (covariates) most highly associated with observed parr densities. We calculated the maximal information coefficient (MIC), using the R package *minerva* (Filosi et al. 2019), to measure the strength of the linear or non-linear association between two variables (Reshef et al. 2011). The MIC value between each of the measured habitat characteristics and parr density was used to inform decisions on which habitat covariates to include in the QRF parr capacity model.

Habitat metrics were first grouped into broad categories that included channel unit, complexity, cover, disturbance, riparian, size, substrate, temperature, water quality, and woody debris. Habitat metrics measuring volume and area were scaled to the wetted area of each site. Within each category, metrics were ranked according to their MIC value (**Figure ??**). Our strategy was to select one or two variables with the highest MIC score within each category so that covariates describe different aspects of rearing habitat (e.g., substrate, temperature, etc.). Additionally, we attempted to avoid covariates that were highly correlated (**Figure ??**) while including covariates that can be directly influenced by restoration actions or have been shown to impact salmonid juvenile density.

2.4 QRF Model Fit

Using the selected habitat covariates (Table 2), we fit a QRF model to predict habitat rearing capacity for spring/summer Chinook salmon parr, during summer months. After constructing a random forest, predictions of the mean response can be made by averaging the predictions of all trees, similar to the expected value predictions from a statistical regression model. However, the individual predictions from each tree, viewed collectively, describe the entire distribution of the predicted response, therefore, the random forest model can be used in the same way as other quantile regression methods to predict any quantile of the response. There were missing values for some habitat data; thus, any site visit with more than three missing covariates was removed from the dataset and the remaining missing habitat values were imputed using the *missForest* R package (Stekhoven and Bühlmann 2012, Stekhoven 2013). We fit the QRF models using the *quantregForest* function from the *quantregForest* package (Meinshausen 2017) in R software (R Core Team 2019), incorporating data from 327 records (paired fish-habitat data) and twelve habitat covariates (27.2 data points per covariate) (Table 2). The 90th quantile of the predicted distribution was used as a proxy for carrying capacity, following the suggestion of Sweka and Mackey (2010). We used the 90th quantile, rather than a higher quantile, to avoid using predictions that are aimed at the very upper tails of observed fish density, where the variability of predictions may be influenced by sample size issues.

After model fitting, the QRF model was then be used to predict capacity using measurements of the habitat covariates that were used to fit the model. In our case, this includes all sites within CHaMP basins in the interior Columbia River basin. For CHaMP sites that have been sampled in multiple years, we first calculated the mean for each habitat metric among years to make predictions. In total, we generated 589 predictions of spring/summer Chinook salmon parr capacity, during summer months, for the following basins: Entiat, Grande Ronde (including Minam), John Day, Lemhi, Methow, South Fork Salmon, Tucannon, Wenatchee and Yankee Fork.

2.5 Extrapolating to Other Sites

Using the QRF model, we predicted habitat capacity for juvenile rearing at all CHaMP sites within the interior Columbia River basin. To predict capacity at larger spatial scales, such as the watershed or population, we developed two extrapolation models (linear and areal capacity) based on globally available attributes (GAA) which were available for the entirety of tributary habitat utilized by a given population. The GAA data used here was taken from the list of GRTS master sample sites that the CHaMP sites were originally selected from (***need citation***). Possible covariates included temperature range, growing degree days, an index of disturbance, the square root of cumulative drainage area, stream power, slope, channel type and watershed (Table 3). The natural log of the CHaMP site capacity predictions was used as the response variable in a multiple linear regression model that incorporated the design weights of the CHaMP sites using the *svyglm* function from the *survey* package (Lumley 2019) in R software (R Core Team 2019). We fit two extrapolation models to each type of fish density, one that included watershed as covariate, for use in predicting capacity within CHaMP watersheds, and one that did not, for predicting everywhere else. We then made predictions of linear and areal capacity at all master sample sites throughout the interior CRB, generally spaced about one kilometer apart. Summaries of extrapolation model fit are shown in Table 4.

To summarize capacity at larger scales, the mean linear capacity (e.g., fish/m) of the master sample points along a particular tributary is multiplied by the length of that tributary. We first restricted the master sample points and lengths of streams to those with the domain of spring/summer Chinook, as defined by StreamNet (http://www.streamnet.org) or using expert opinion from local biologists. The capacities of various tributaries could then be summed to estimate capacity at almost any spatial scale. However, for visualization, predictions of areal capacity (fish/m^2) were used.

2.6 Model Validation

Spawner-recruit data from several watersheds within the interior CRB were compiled to validate the extrapolated QRF estimates of spring/summer Chinook salmon parr capacity. Some watersheds had direct estimates of parr, while some had estimates of smolts and pre-smolts from rotary screw traps. For the latter, estimates of parr were calculated using estimates of over-winter survival. A series of spawner-recruit functions were then fit to this data including Beverton-Hold, Ricker, and hockey stick. Estimates of capacity were made from each of these spawner-recruit curves and compared with QRF estimates of capacity (Table 5, Figure 4).

3 Results

3.1 Habitat Covariate Selection

We categorized 164 habitat measurements collected using the CHaMP habitat protocol (CHaMP (Columbia Habitat Monitoring Program) 2016) into eleven habitat categories, and for each habitat covariate an MIC value was calculated based on the strength of association between the habitat covariate and the response variable, parr density (fish/m). We chose the following twelve CHaMP habitat covariates to fit the QRF model: wetted width:depth ratio, channel unit frequency, standard deviation of the wetted depth, braidedness, total fish cover, percent riparian understory, wetted width, observed discharge, percent fines less than 6mm, lower quantile of substrate size (D16), average august temperature and the frequency of large wood in the wetted channel (Table 2).

3.2 QRF Model

A QRF model was fit using those metrics and the *quantregForest* package (Meinshausen 2006) in R (R Core Team 2019) and the 90th quantile of the predicted distribution was used as a proxy for carrying capacity. After model fit, we examined the relative importance of each habitat covariate included in the model (Figure 2). Moreover, QRF models allow one to visually examine the marginal effect of each habitat covariate on the quantile of interest using partial dependence plots (PDP). These plots show the marginal effect of changing a single habitat covariate while maintaining all other covariates at their mean values (Figure 3). However, given that many habitat metrics are somewhat correlated, these marginal effects are not independent of one another

making the interpretation of partial dependence plots less straightforward. After model fitting, the QRF model was used to predict habitat capacity at all CHaMP sites within the interior Columbia River basin.

3.3 Extrapolating to Other Sites

The coefficients for the extrapolation model can be found in Tbale 3.

3.4 Model Validation

Estimates of Chinook salmon parr capacity from the QRF and extrapolation models were comparable to independent estimates from spawner-recruit data (Table 5, Figure 4). QRF estimates had overlapping confidence intervals with one or more of the Beverton-Holt, Ricker, or hockey stick model estimates in each of the seven locations where comparisons were possible (Figure 4). Possible additional uncertainty was not accounted for in estimates of spawners-per-redd or spawners-per-parr, which would increase the confidence intervals around spawner-recruit estimates and the overlap among estimates. Correlations between parr capacity estimates from the QRF model and spawner-recruit models ranged from 0.905 (Beverton-Holt) to 0.965 (hockey stick). This favorable comparison provides strong validation, as the spawner-recruit estimates of Chinook salmon parr capacity were developed from completely independent datasets and using entirely different methods.

4 Discussion

4.1 A Tool To Estimate Habitat Capacity

In this study, we developed a novel approach to estimate the capacity of habitat to support Chinook salmon parr during summer months in wadeable streams. Our model can be used to quantify juvenile rearing capacity in Chinook salmon watersheds or populations and, in turn, quantify the magnitude of tributary habitat rehabilitation necessary to support Endangered Species Act (ESA) delisting. The QRF and extrapolation models presented here provide useful tools towards the prioritization, implementation, and evaluation of habitat restoration actions to recover depleted salmon populations. Moreover, the models presented here can be applied to multiple stages within the life cycle (e.g., parr, smolt, adult). Estimates of habitat carrying capacity for multiple life stages will allow biologists and managers to identify what life stage specific habitat is limiting. For example, QRF models and associated extrapolation models may demonstrate that habitat for a given population is sufficient to support adult spawning required for ESA delisting, but that juvenile capacity may not be sufficient to support those levels of adult spawning. In such a case, habitat restoration actions may be most cost-effectively and sustainably directed towards juvenile rearing habitat. Models to estimate habitat carrying capacity for multiple life stages will help to better direct habitat restoration actions.

4.2 What Watershed-scale Capacity Models Are Currently Available?

Historically, fish-habitat relationship models have focused on species distribution or average abundance/density rather than directly addressing carrying capacity. The QRF approach employed here allows for the analysis of multiple correlated habitat metrics, often with non-linear relationships to fish density, in order to assess fish habitat relationships for higher quantiles that are assumed to represent carrying capacity.

4.3 Biological Expectations from QRF Model

The results of the QRF parr capacity model for spring/summer Chinook salmon meet many biological expectations.

4.4 Extrapolation Model

In addition to site-specific estimates of carrying capacity derived from paired fish-habitat data, our extrapolation model allows for analyses at larger extents, such as watershed and population level scales. This is an efficient technique to leverage existing relationships for meaningful management decisions.

While the Columbia River Basin is an excellent candidate for this type of extrapolation given the wealth of CHaMP data, application of this technique to areas with limited habitat data may be considerably more challenging. On-the-ground habitat data collection is costly, time-consuming, and often not reproducible; thereby limiting the effectiveness of this type of extrapolation. Additionally, the extrapolation model relies on Global Available Attribute (GAA) data which ....

Despite these considerations, we were able to leverage extensive datasets to produce carrying capacity estimates throughout the CRB that significantly increase or habitat restoration prioritization in a cost-effective manner.

4.5 The Future: Improving Habitat Data

Given the cost/extent of data necessary for QRF extrapolation in watersheds outside of the CRB, there is a pressing need to develop new tools for habitat analyses. Unmanned Aerial Systems (UAS or drone, commonly) are gaining popularity in wildlife and ecosystem monitoring for their ease of use, safety, accessibility, and cost-efficiency (Jones IV et al. 2006, Chabot and Bird 2015). UAS produce high-resolution, permanent data at a fraction of the cost of on-the-ground habitat sampling. Advances in imaging techniques (e.g., multispectral imaging) and post processing (e.g., automation of data collection from imagery) are already demonstrating increase in the efficiency and accuracy of data collection (Whitehead and Hugenholtz 2014, LeCun et al. 2015, Weinstein 2018).

Development of a standardized protocol to incorporate remotely sensed data (LiDAR, aerial imagery) into the collection of habitat metrics would greatly improve the broadscale application of QRF. Rapid advances drone technology further improves upon traditional habitat data collection by leveraging 1) sub-meter global navigation satellite system (GNSS) receivers; 2) cost-effective drone imagery collection, image stitching, and photogrammetry; and 3) semi-automated data post-processing. All data collection efforts would be georeferenced and topologically compatible to increase repeatability of methods and data collection locations; a primary criticism of previous CHaMP survey efforts. The implementation of such a protocol would circumvent the need to extrapolate by collecting data for individual channel units in a rapid manner using remote sensing technologies, thereby reducing labor, providing a cost-effective tool for habitat data collection supporting status and trend evaluation and model products to better inform habitat restoration prioritization and planning.

4.6 Conclusions and Next Steps

5 Acknowledgements

Model development efforts have been funded by the Bonneville Power Administration through projects 2003-017-00 and 2011-006-00 and by the Bureau of Reclamation and Idaho Office of Species Conservation through contract BOR002 16. Fish sampling work in the Lemhi River was also funded through the Idaho Office of Species Conservation through the Pacific Coast Salmon Recovery Fund. Special thanks to staff from the following agencies for providing data to calculate Chinook salmon parr abundance and density estimates: Columbia River Inter-Tribal Fish Commission, Oregon Department of Fish and Wildlife, and U.S. Fish and Wildlife Service. The models in this study were improved by conversations with Eric Buhle.

6 References

Braun, D. C., and J. D. Reynolds. 2011. Relationships between habitat characteristics and breeding population densities in sockeye salmon (oncorhynchus nerka). Canadian Journal of Fisheries and Aquatic Sciences 68:758–767.

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

Bryant, M., and R. Woodsmith. 2009. The response of salmon populations to geomorphic measurements at three scales. North American Journal of Fisheries Management 29:549–559.

Cade, B. S., and B. R. Noon. 2003. A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment 1:412–420.

Carle, F. L., and M. R. Strub. 1978. A new method for estimating population size from removal data. Biometrics:621–630.

Chabot, D., and D. M. Bird. 2015. Wildlife research and management methods in the 21st century: Where do unmanned aircraft fit in? Journal of Unmanned Vehicle Systems 3:137–155.

CHaMP (Columbia Habitat Monitoring Program). 2016. Scientific protocol for salmonid habitat surveys within the columbia habitat monitoring program. Bonneville Power Administration Portland, Oregon, USA.

Chapman, D. G. 1951. Some properties of the hypergeometric distribution with applications to zoological sample censuses. University of California Press.

Compton, T. J., M. A. Morrison, J. R. Leathwick, and G. D. Carbines. 2012. Ontogenetic habitat associations of a demersal fish species, pagrus auratus, identified using boosted regression trees. Marine Ecology Progress Series 462:219–230.

Cramer, S. P., and N. K. Ackerman. 2009. Linking stream carrying capacity for salmonids to habitat features. Pages 225–254 *in* American fisheries society symposium.

Dunham, J. B., B. S. Cade, and J. W. Terrell. 2002. Influences of spatial and temporal variation on fish-habitat relationships defined by regression quantiles. Transactions of the American Fisheries Society 131:86–98.

Eastwood, P. D., G. J. Meaden, A. Carpentier, and S. I. Rogers. 2003. Estimating limits to the spatial extent and suitability of sole (solea solea) nursery grounds in the dover strait. Journal of Sea Research 50:151–165.

Fausch, K., C. Hawkes, and M. Parsons. 1988. Models that predict standing crop of stream fish from habitat variables: 1950-85. Notes.

Filosi, M., R. Visintainer, and D. Albanese. 2019. Minerva: Maximal information-based nonparametric exploration for variable analysis.

Good, T. P., R. S. Waples, and P. B. Adams. 2005. Updated status of federally listed ESUs of west coast salmon and steelhead.

Hedger, R. D., E. De Eyto, M. Dillane, O. H. Diserud, K. Hindar, P. McGinnity, R. Poole, and G. Rogan. 2013. Improving abundance estimates from electrofishing removal sampling. Fisheries Research 137:104–115.

Jones IV, G. P., L. G. Pearlstine, and H. F. Percival. 2006. An assessment of small unmanned aerial vehicles for wildlife research. Wildlife society bulletin 34:750–758.

Knudby, A., A. Brenning, and E. LeDrew. 2010. New approaches to modelling fish-habitat relationships. Ecological Modelling 221:503–511.

Koenker, R., and G. Bassett Jr. 1978. Regression quantiles. Econometrica: journal of the Econometric Society:33–50.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. Nature 521:436–444.

Lumley, T. 2019. Survey: Analysis of complex survey samples.

Meinshausen, N. 2006. Quantile regression forests. Journal of Machine Learning Research 7:983–999.

Meinshausen, N. 2017. QuantregForest: Quantile regression forests.

Mossop, B., and M. J. Bradford. 2006. Using thalweg profiling to assess and monitor juvenile salmon (oncorhynchus spp.) habitat in small streams. Canadian Journal of Fisheries and Aquatic Sciences 63:1515–1525.

Nickelson, T. E., J. D. Rodgers, S. L. Johnson, and M. F. Solazzi. 1992. Seasonal changes in habitat use by juvenile coho salmon (oncorhynchus kisutch) in oregon coastal streams. Canadian Journal of Fisheries and Aquatic Sciences 49:783–789.

Ogle, D., P. Wheeler, and A. Dinno. 2019. FSA: Simple fisheries stock assessment methods.

Pess, G. R., D. R. Montgomery, E. A. Steel, R. E. Bilby, B. E. Feist, and H. M. Greenberg. 2002. Landscape characteristics, land use, and coho salmon (oncorhynchus kisutch) abundance, snohomish river, wash., usa. Canadian Journal of Fisheries and Aquatic Sciences 59:613–623.

Pittman, S., B. Costa, and T. Battista. 2009. Using lidar bathymetry and boosted regression trees to predict the diversity and abundance of fish and corals. Journal of Coastal Research 53:27–38.

Prasad, A., L. 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. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reshef, D. N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. 2011. Detecting novel associations in large data sets. Science 334:1518–1524.

Rivest, L.-P., and S. Baillargeon. 2019. Rcapture: Loglinear models for capture-recapture experiments.

Robson, D., and H. Regier. 1964. Sample size in petersen mark–recapture experiments. Transactions of the American Fisheries Society 93:215–226.

Seber, G. 2002. The estimation of animal abundance and related parameters. Blackburn Press Caldwell, New Jersey.

Sharma, R., and R. Hilborn. 2001. Empirical relationships between watershed characteristics and coho salmon (oncorhynchus kisutch) smolt abundance in 14 western washington streams. Canadian Journal of Fisheries and Aquatic Sciences 58:1453–1463.

Stekhoven, D. J. 2013. MissForest: Nonparametric missing value imputation using random forest.

Stekhoven, D. J., and P. Bühlmann. 2012. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28:112–118.

Stevens Jr, D., and A. R. Olsen. 1999. Spatially restricted surveys over time for aquatic resources. Journal of Agricultural, Biological, and Environmental Statistics:415–428.

Stevens Jr, D., and A. R. Olsen. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99:262–278.

Sweka, J. A., and G. Mackey. 2010. A functional relationship between watershed size and atlantic salmon parr density. Journal of Fish and Wildlife Management 1:3–10.

Terrell, J. W., B. S. Cade, J. Carpenter, and J. M. Thompson. 1996. Modeling stream fish habitat limitations from wedge-shaped patterns of variation in standing stock. Transactions of the American Fisheries Society 125:104–117.

Volk, C., N. Bouwes, C. Jordan, J. Wheaton, P. Nelle, C. Beasley, S. Walker, B. Bouwes, M. Nahorniak, A. Hill, J. Heitke, K. Whitehead, S. Bangen, and E. Ward. 2017. Integrated status and effectiveness monitoring program (ISEMP) and columbia habitat monitoring program (CHaMP) 2016 annual combined technical report. Bonneville Power Administration.

Weinstein, B. G. 2018. A computer vision for animal ecology. Journal of Animal Ecology 87:533–545.

Whitehead, K., and C. H. Hugenholtz. 2014. Remote sensing of the environment with small unmanned aircraft systems (uass), part 1: A review of progress and challenges. Journal of Unmanned Vehicle Systems 2:69–85.

7 Tables

Table 1: The number of unique sites, by watershed, with paired fish-habitat data used to populate the spring/summer Chinook salmon parr capacity QRF model

Watershed	n Sites	Percent
Entiat	61	18.7%
John Day	75	22.9%
Lemhi	33	10.1%
Minam	20	6.1%
South Fork Salmon	30	9.2%
Upper Grande Ronde	86	26.3%
Wenatchee	22	6.7%
Total	327	100.0%

Tabl
the
capa
relat

Rank	Metric	Metric Category	Description
1	Wetted Width Integrated	Size	Average width of the wetted polygon for a site.
2	Discharge	Size	The sum of station discharge across all stations. Station d
3	Avg. August Temperature	Temperature	Average predicted daily August temperature from NorWes
4	Wetted Width To Depth Ratio Avg	Complexity	Average width to depth ratio of the wetted channel measu
5	Substrate < 6mm	Substrate	Average percentage of pool tail substrates comprised of se
6	Large Wood Frequency: Wetted	Wood	Number of large wood pieces per 100 meters within the we
7	Fish Cover: Total	Cover	Percent of wetted area with the following types of cover: a
8	Channel Unit Frequency	ChannelUnit	Number of channel units per 100 meters.
9	Wetted Depth SD	Complexity	Standard Deviation of water depths within the wetted cha
10	Riparian Cover: Understory	Riparian	Percent of understory vegetation cover.
11	Substrate: D16	Substrate	Diameter of the 16th percentile particle derived from pebb
12	Wetted Channel Braidedness	Complexity	Ratio of the total length of the wetted mainstem channel p

Table 3: Globally available attribute (GAA) habitat covariates used to extrapolate quantile regression forest (QRF) model predictions of spring/summer Chinook parr capacity to a larger scale (e.g., watershed, population), with their coefficients and standard errors.

Covariate	Name	Scale	Estimate	Std. Error
TRange	Temeprature Range	Reach (2 km)	-0.014	0.079
Elev_M	Elevation	Site (300 m)	-0.187	0.158
CHaMPsheds	CHaMP Watershed	Reach (2 km)	-	-
NatPrin1	Natural Class PCA 1	Watershed (HUC12)	-0.112	0.072
DistPrin1	Disturbance Class PCA 1	Watershed (HUC12)	-0.033	0.064
SrtCumDrn	Drainage Area (sqrt)	Reach (2 km)	-0.178	0.083
StrmPwr	Stream Power	Reach (2 km)	0.074	0.030
Slp_NHD_v1	Slope	Reach (2 km)	-0.540	0.114
Channel_Type	Channel Type	Site (300 m)	-	-

Covariate	Name	Scale	Estimate	Std. Error
WIDE_BF	Bankfull Width - modeled	Site (300 m)	0.210	0.106
S2_02_11	NorWeST Avg. Aug. Temp	Reach (2 km)	-0.107	0.120

Table 4: Summary of extrapolation model fits, split by each of the responses and whether the extrapolation model used CHaMP watershed as a covariate or not.

Model	Response	r^2	Adjusted r^2
CHaMP	fish/m	0.495	0.469
non-CHaMP	fish/m	0.371	0.350
CHaMP	fish/m^2	0.373	0.341
non-CHaMP	fish/m^2	0.214	0.188

Table 5: Estimates of parr capacity from both spawner-recruit data (Beverton-Holt, Ricker, hockey stick) and from extrapolated estimates of parr capacity from the quantile regression forest (QRF) model. Numbers in parentheses are coefficients of variation.

Population	n Yrs	Adult Data	Parr Data	Beverton Holt	Ricker	Hockey Stick
Catherine Creek	20	Spawners	RST	135,387 (0.269)	103,021 (0.141)	99,921 (0.21)
Chiwawa R.	20	Spawners	Parr Surveys	248,586 (0.24)	166,139 (0.148)	174,216 (0.184)
Hayden Creek	7	Spawners	RST	58,394 (0.244)	65,958 (0.195)	48,351 (0.174)
Lostine River	17	Redds	RST	196,259 (0.24)	146,982 (0.159)	144,415 (0.201)
Minam River	14	Spawners	RST	1,309,223 (2.18)	484,810 (1.444)	662,806 (1.726)
South Fork Salmon River	17	Redds	RST	87,260 (0.407)	62,456 (0.265)	64,653 (0.317)
Tucannon River	27	Redds	RST	4,791,131 (13.016)	1,234,653 (8.566)	NA (NA)
Upper Grande Ronde River	8	Spawners	RST	171,607 (0.388)	168,137 (0.298)	127,052 (0.317)
Upper Lemhi	7	Spawners	RST	333,229 (0.322)	229,635 (0.212)	242,636 (0.252)

8 Figures

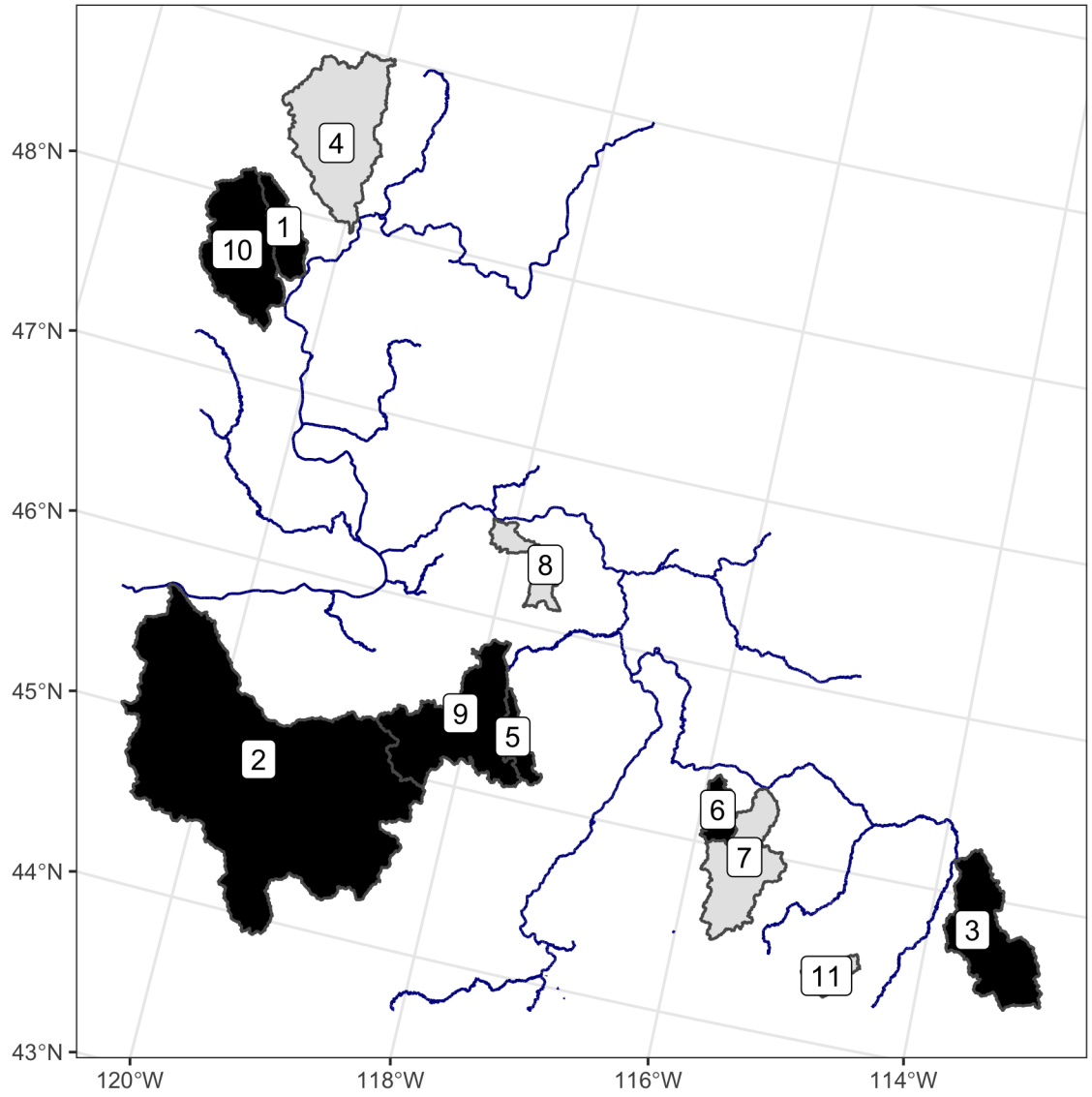

Figure 1: Watersheds with CHaMP habitat data. Watersheds in black also contain paired fish data. Watershed names are: 1 - Entiat, 2 - John Day, 3 - Lemhi, 4 - Methow, 5 - Minam, 6 - Secesh, 7 - South Fork Salmon, 8 - Tucannon, 9 - Upper Grande Ronde, 10 - Wenatchee, 11 - Yankee Fork.

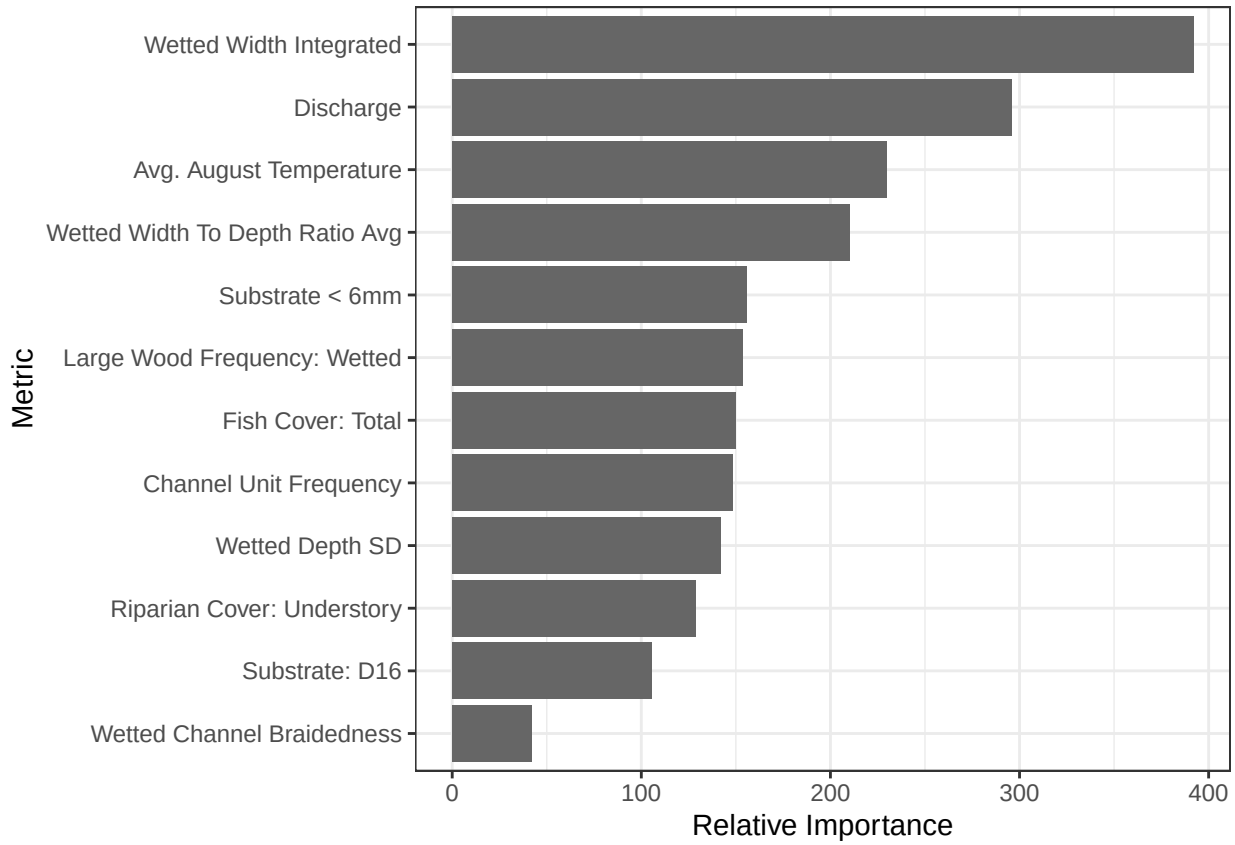

Figure 2: Relative importance of each habitat covariate included in the quantile regression forest (QRF) model to predict habitat capacity, during summer months, for spring/summer Chinook salmon parr

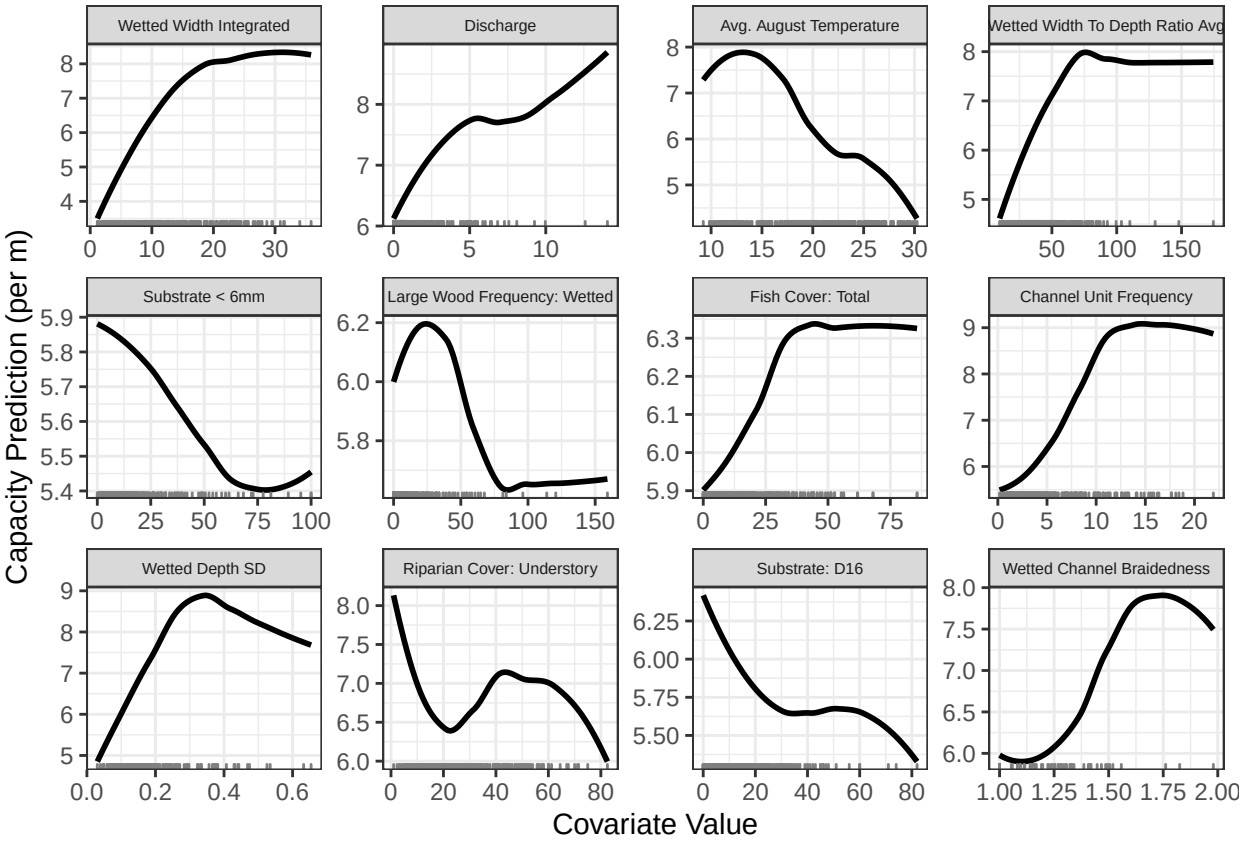

Figure 3: Partial dependence plots for the spring/summer Chinook salmon parr capacity quantile regression forest (QRF) model, depicting how parr capacity shifts as each habitat metric changes, assuming all other habitat metrics remain at their mean values. Tick marks along the X-axis depict observed values.

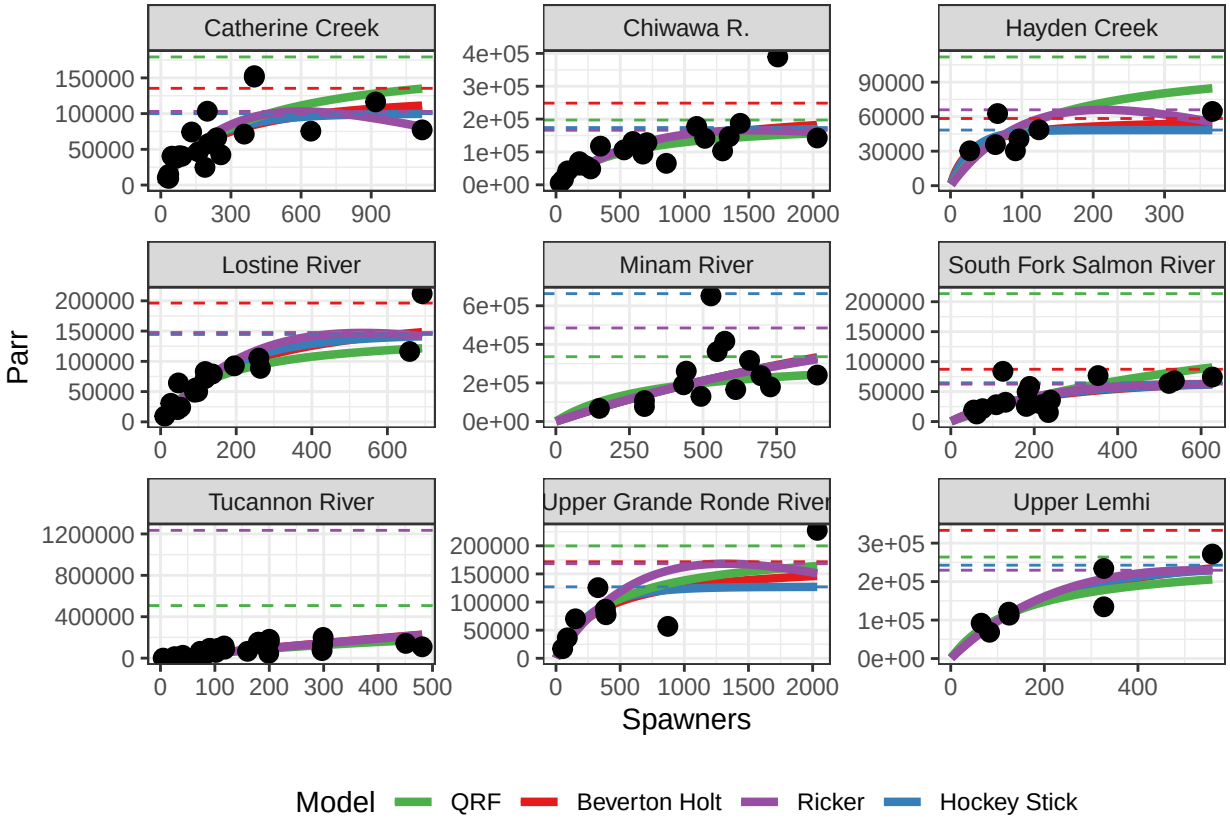

Figure 4: Spawner-recruit data from seven watersheds. Solid lines are the spawner-recruit curve, dashed lines are the estimated capacity, and shaded polygons depict the 95% confidence intervals of capacity. Blue corresponds to Beverton-Holt models, green to Ricker models, yellow to hockey stick models, and red to QRF estimates. The QRF solid curve is a Beverton-Holt model with the capacity parameter fixed to the QRF estimate of capacity.

8.0.1 Colophon

This report was generated on 2020-02-04 15:22:29 using the following computational environment and dependencies:

```
#> - Session info ---------------------------------------------------------------
#>  setting  value
#>  version  R version 3.6.1 (2019-07-05)
#>  os       macOS Mojave 10.14.6
#>  system   x86_64, darwin15.6.0
#>  ui       X11
#>  language (EN)
#>  collate  en_US.UTF-8
#>  ctype    en_US.UTF-8
#>  tz       America/Los_Angeles
#>  date     2020-02-04
#>
#> - Packages -------------------------------------------------------------------
#>  package        * version    date       lib source
#>  assertthat       0.2.1      2019-03-21 [2] CRAN (R 3.6.0)
#>  backports        1.1.5      2019-10-02 [2] CRAN (R 3.6.0)
#>  bookdown         0.17       2020-01-11 [1] CRAN (R 3.6.0)
#>  broom            0.5.3      2019-12-14 [1] CRAN (R 3.6.0)
#>  callr            3.4.0      2019-12-09 [2] CRAN (R 3.6.1)
#>  cellranger       1.1.0      2016-07-27 [2] CRAN (R 3.6.0)
#>  class            7.3-15     2019-01-01 [2] CRAN (R 3.6.1)
#>  classInt         0.4-2      2019-10-17 [1] CRAN (R 3.6.0)
#>  cli              2.0.1      2020-01-08 [1] CRAN (R 3.6.0)
#>  colorspace       1.4-1      2019-03-18 [2] CRAN (R 3.6.0)
#>  crayon           1.3.4      2017-09-16 [2] CRAN (R 3.6.0)
#>  DBI              1.0.0      2018-05-02 [2] CRAN (R 3.6.0)
#>  dbplyr           1.4.2      2019-06-17 [2] CRAN (R 3.6.0)
#>  desc             1.2.0      2018-05-01 [1] CRAN (R 3.6.0)
#>  devtools         2.2.1      2019-09-24 [1] CRAN (R 3.6.0)
#>  digest           0.6.23     2019-11-23 [2] CRAN (R 3.6.0)
#>  dplyr          * 0.8.3      2019-07-04 [2] CRAN (R 3.6.0)
#>  e1071            1.7-3      2019-11-26 [1] CRAN (R 3.6.0)
#>  ellipsis         0.3.0      2019-09-20 [2] CRAN (R 3.6.0)
#>  english        * 1.2-5      2020-01-26 [1] CRAN (R 3.6.1)
#>  evaluate         0.14       2019-05-28 [2] CRAN (R 3.6.0)
#>  fansi            0.4.1      2020-01-08 [1] CRAN (R 3.6.0)
#>  farver           2.0.3      2020-01-16 [1] CRAN (R 3.6.0)
#>  forcats        * 0.4.0      2019-02-17 [2] CRAN (R 3.6.0)
#>  foreign          0.8-71     2018-07-20 [2] CRAN (R 3.6.1)
#>  fs               1.3.1      2019-05-06 [2] CRAN (R 3.6.0)
#>  generics         0.0.2      2018-11-29 [2] CRAN (R 3.6.0)
#>  ggplot2        * 3.2.1      2019-08-10 [2] CRAN (R 3.6.0)
#>  glue             1.3.1      2019-03-12 [2] CRAN (R 3.6.0)
#>  gtable           0.3.0      2019-03-25 [2] CRAN (R 3.6.0)
#>  haven            2.2.0      2019-11-08 [2] CRAN (R 3.6.0)
#>  highr            0.8        2019-03-20 [2] CRAN (R 3.6.0)
#>  hms              0.5.2      2019-10-30 [2] CRAN (R 3.6.0)
#>  htmltools        0.4.0      2019-10-04 [2] CRAN (R 3.6.0)
#>  httr             1.4.1      2019-08-05 [2] CRAN (R 3.6.0)
#>  janitor        * 1.2.0      2019-04-21 [1] CRAN (R 3.6.0)
```

```
#>  jsonlite         1.6        2018-12-07 [2] CRAN (R 3.6.0)
#>  KernSmooth       2.23-15    2015-06-29 [2] CRAN (R 3.6.1)
#>  knitr          * 1.27       2020-01-16 [1] CRAN (R 3.6.0)
#>  labeling         0.3        2014-08-23 [2] CRAN (R 3.6.0)
#>  lattice          0.20-38    2018-11-04 [2] CRAN (R 3.6.1)
#>  lazyeval         0.2.2      2019-03-15 [2] CRAN (R 3.6.0)
#>  lifecycle        0.1.0      2019-08-01 [2] CRAN (R 3.6.0)
#>  lubridate        1.7.4      2018-04-11 [2] CRAN (R 3.6.0)
#>  magrittr       * 1.5        2014-11-22 [2] CRAN (R 3.6.0)
#>  maptools         0.9-5      2019-02-18 [1] CRAN (R 3.6.1)
#>  Matrix         * 1.2-17     2019-03-22 [2] CRAN (R 3.6.1)
#>  memoise          1.1.0      2017-04-21 [1] CRAN (R 3.6.0)
#>  mitools          2.4        2019-04-26 [1] CRAN (R 3.6.0)
#>  modelr           0.1.5      2019-08-08 [2] CRAN (R 3.6.0)
#>  munsell          0.5.0      2018-06-12 [2] CRAN (R 3.6.0)
#>  nlme             3.1-142    2019-11-07 [1] CRAN (R 3.6.0)
#>  pander         * 0.6.3      2018-11-06 [1] CRAN (R 3.6.0)
#>  pillar           1.4.3      2019-12-20 [1] CRAN (R 3.6.0)
#>  pkgbuild         1.0.6      2019-10-09 [1] CRAN (R 3.6.0)
#>  pkgconfig        2.0.3      2019-09-22 [2] CRAN (R 3.6.0)
#>  pkgload          1.0.2      2018-10-29 [1] CRAN (R 3.6.0)
#>  png              0.1-7      2013-12-03 [1] CRAN (R 3.6.0)
#>  prettyunits      1.1.0      2020-01-09 [1] CRAN (R 3.6.0)
#>  processx         3.4.1      2019-07-18 [2] CRAN (R 3.6.0)
#>  ps               1.3.0      2018-12-21 [2] CRAN (R 3.6.0)
#>  purrr          * 0.3.3      2019-10-18 [2] CRAN (R 3.6.0)
#>  QRFpaper       * 0.0.0.9000 2020-01-24 [1] local
#>  quantregForest   1.3-7      2017-12-19 [2] CRAN (R 3.6.0)
#>  R6               2.4.1      2019-11-12 [2] CRAN (R 3.6.0)
#>  randomForest     4.6-14     2018-03-25 [2] CRAN (R 3.6.0)
#>  RColorBrewer     1.1-2      2014-12-07 [2] CRAN (R 3.6.0)
#>  Rcpp             1.0.3      2019-11-08 [2] CRAN (R 3.6.0)
#>  readr          * 1.3.1      2018-12-21 [2] CRAN (R 3.6.0)
#>  readxl           1.3.1      2019-03-13 [2] CRAN (R 3.6.0)
#>  remotes          2.1.0      2019-06-24 [1] CRAN (R 3.6.0)
#>  reprex           0.3.0      2019-05-16 [2] CRAN (R 3.6.0)
#>  rlang            0.4.2      2019-11-23 [2] CRAN (R 3.6.0)
#>  rmarkdown        2.1        2020-01-20 [1] CRAN (R 3.6.1)
#>  rprojroot        1.3-2      2018-01-03 [1] CRAN (R 3.6.0)
#>  rstudioapi       0.10       2019-03-19 [2] CRAN (R 3.6.0)
#>  rvest            0.3.5      2019-11-08 [2] CRAN (R 3.6.0)
#>  scales           1.1.0      2019-11-18 [2] CRAN (R 3.6.0)
#>  sessioninfo      1.1.1      2018-11-05 [1] CRAN (R 3.6.0)
#>  sf             * 0.8-0      2019-09-17 [1] CRAN (R 3.6.0)
#>  sp               1.3-2      2019-11-07 [1] CRAN (R 3.6.0)
#>  stringi          1.4.5      2020-01-11 [1] CRAN (R 3.6.0)
#>  stringr        * 1.4.0      2019-02-10 [2] CRAN (R 3.6.0)
#>  survey         * 3.36       2019-04-27 [1] CRAN (R 3.6.0)
#>  survival       * 2.44-1.1   2019-04-01 [2] CRAN (R 3.6.1)
#>  testthat         2.3.1      2019-12-01 [1] CRAN (R 3.6.0)
#>  tibble         * 2.1.3      2019-06-06 [2] CRAN (R 3.6.0)
#>  tidyr          * 1.0.0      2019-09-11 [2] CRAN (R 3.6.0)
#>  tidyselect       0.2.5      2018-10-11 [2] CRAN (R 3.6.0)
#>  tidyverse      * 1.3.0      2019-11-21 [2] CRAN (R 3.6.0)
```

```
#>  units            0.6-5      2019-10-08 [1] CRAN (R 3.6.0)
#>  usethis          1.5.1      2019-07-04 [1] CRAN (R 3.6.0)
#>  vctrs            0.2.1      2019-12-17 [1] CRAN (R 3.6.0)
#>  withr            2.1.2      2018-03-15 [2] CRAN (R 3.6.0)
#>  xfun             0.12       2020-01-13 [1] CRAN (R 3.6.0)
#>  xml2             1.2.2      2019-08-09 [2] CRAN (R 3.6.0)
#>  yaml             2.2.0      2018-07-25 [2] CRAN (R 3.6.0)
#>  zeallot          0.1.0      2018-01-28 [2] CRAN (R 3.6.0)
#>
#> [1] /Users/seek/Library/R/3.6/library
#> [2] /Library/Frameworks/R.framework/Versions/3.6/Resources/library
```

The current Git commit details are:

```
#> Local:    master /Users/seek/Documents/GitProjects/MyProjects/QRFpaper
#> Remote:   master @ origin (git@github.com:KevinSee/QRFpaper.git)
#> Head:     [89b21ef] 2020-02-04: Merge branch 'master' of github.com:KevinSee/QRFpaper
```