**Life-Stage Specific Habitat Capacity for Anadromous Salmonids using Quantile Regression Forest Models**

Kevin See1,2, Michael W. Ackerman1, Richard A. Carmichael1, Chris Beasley1

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

2Corresponding author: [kevin.see@biomark.com](mailto:kevin.see@biomark.com)

**Running Head:** QRF to Estimate Habitat Capacity

**Keywords:**

**Highlights:**

**Suggested Reviewers:**

# Abstract

# 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, hydrosystem 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 cost-effectively and sustainably reduce mortality. Recently released “delisting” criteria (NOAA 2016) identified adult escapement targets at the population scale, providing a quantitative metric useful for evaluating the magnitude of survival improvements (across life stages) required. These abundance targets provide a benchmark against which habitat rehabilitation actions can be measured. In this paper, we describe a novel approach for estimating life-stage specific habitat capacity that can be used to quantitatively identify the magnitude of tributary habitat restoration necessary to support Endangered Species Act (ESA) delisting. For perhaps the first time, the necessity of tributary habitat restoration actions can be demonstrated and the magnitude of required change can be placed in context with the other “H’s”.

In combination with the other three “H’s” described above, efforts to recover depleted salmon populations are replete with actions to rehabilitate habitat used during the freshwater life-stages (Good et al. 2005). Specifically, restoring salmonid carrying capacity through tributary rehabilitation is a key component of recovery plans for salmon and steelhead (*Oncorhynchus mykiss*) in the Pacific Northwest, USA. Efforts to recover population productivity have included increasing tributary habitat quantity and quality for spawning adults and rearing juveniles. The goal of habitat restoration is often to restore habitat quantity/quality to historic levels to increase the capacity to support spawning adults and rearing juveniles.

Carrying capacity is the maximal abundance or load the habitat can support for a given species and life-stage given current resources (citation). Why do we think carrying capacity is an important metric for measuring recover? Estimating habitat carrying capacity, both historic and contemporary, as well as identifying important habitat characteristics that influence capacity for various life-stages of Pacific salmon is an ongoing 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. Such methods will provide tools to direct and prioritize tributary habitat 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 (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). Byrant 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 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, Knuby 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 and it often fails to account for habitat covariates (Cramer and Ackerman, 2009).

Historically, fisheries research and management has relied on the assumption that higher observed relative fish densities within a given life stage are a function of habitat quantity and quality. However, observed fish density can be a poor predictor of habitat capacity owing to a paucity of individuals, which is further confounded by unmeasured biotic or abiotic variables. To address this, we developed a quantile regression forest (QRF; Meinshausen 2006) model to estimate the carrying capacity of wadeable streams to support rearing juvenile Pacific salmon. 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. QRF models combine the flexibility of random forest models (Breiman 2001) with the ability of quantile regression to extract relationships from any quantile of the data (Cade and Noon 2003). Random forest models capture non-linear relationships between response (fish) and predictor (habitat) variables while naturally incorporating interactions between predictor variables, common features of ecological datasets (Liaw and Wiener 2002). Additionally, random forest models have been shown to outperform more standard parametric models in predicting fish-habitat relationships in other contexts (Knudby et al. 2010) and can work with untransformed data while being robust to outliers (Breiman 2001; Prassad et al. 2006). Meanwhile, q to estimate the effect of limiting factors

In this study, we developed a QRF model to better elicit fish-habitat relationships and predict habitat rearing capacity at the site scale for juvenile stream-type Chinook salmon (*Oncorhynchus tshawytscha*; hereafter Chinook salmon) using paired fish and habitat data. Fish and habitat data used to populate the QRF model are available from eight watersheds within the interior Columbia River Basin, Pacific Northwest, USA. Fish survey data 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 and the Snake River spring/summer-run is listed as threatened. Habitat data are from the Columbia Habitat Monitoring Program (CHaMP; <https://www.champmonitoring.org>). Fish and habitat data were paired at CHaMP sites (200 – 500 m) where fish survey data were available. Specifically, we focus on the summer parr (juvenile) life-stage of Chinook salmon and we were interested in predicting the capacity of contemporary habitat to support Chinook salmon parr through the summer months following emergence. Importantly, the QRF model places no constraints on possible fish-habitat relationships; instead, relationships are estimated from the data regardless of being positive, negative, linear, non-linear, etc. (Meinshausen 2006). Based on the observed fish-habitat relationships, we can then predict capacity using measurements of the habitat characteristics that were used to populate the model. In summary, our objectives were 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
5. Validate estimates of carrying capacity from QRF across multiple watersheds using independent estimates of capacity (e.g., spawner-recruit relationships)

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.

# 2. Methods

## 2.1 Study Site

Fish and habitat data used in the QRF model were collected from nine watersheds within the interior Columbia River Basin, Pacific Northwest, USA (Figure 1). The Columbia River Basin covers more than 668,000 km2; draining substantial portions of Idaho, Oregon, and Washington, smaller portions of 4 additional states, as well as the southeastern portion of British Columbia. Data from the following nine watersheds within the Columbia River Basin were used in this study: Entiat, Grande Ronde (upper), John Day, Lemhi, Methow, Minam (tributary of Grande Ronde), Secesh, Tucannon, and Wenatchee (Table 1). Juvenile density and abundance data collected at CHaMP survey sites were graciously provided by a number of projects included the Integrated Status and Effectiveness Monitoring Program (ISEMP).

## 2.2 Data

The habitat data were collected by the CHaMP (CHaMP; ISEMP/CHaMP 2017) and were downloaded from <https://www.champmonitoring.org>. CHaMP sites are 200 m to 500 m reaches (within wadeable streams) based on a spatially balanced Generalized Random Tesselation Stratified sample selection algorithm (Stevens and Olsen 1999, 2004). Habitat data within CHaMP sites were collected using the CHaMP protocol (CHaMP 2016) 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.

Chinook salmon parr were sampled during the summer low-flow season at many of the same sites that were surveyed for habitat using the CHaMP protocol. Fish survey methods included mark-recapture, three-pass removal, two-pass removal, and single-pass electrofishing, as well as snorkeling. Survey data were used to estimate Chinook salmon parr abundance at all sites where data were available. 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). 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). These abundance estimates were made using the *removal* function from the *FSA* package (Ogle 2017) or the *closed.bc* function from the *Rcapture* package (Rivest and Baillargeon 2014) in R software (R Core Team 2015). Abundance was estimated from snorkel counts by using a correction factor derived from sites where paired snorkel-electrofishing data were available. 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 model that was developed separately for each basin and sampling entity. The ratio estimator was based on a binomial generalized linear model that was developed separately for each basin and each sampling agency. Possible covariates included the number of fish captured on the first pass, year, site length, and Julian day. After fitting all possible models with those covariates to data with valid abundance estimates, the model with the lowest AICc for each basin and sampling agency was chosen and used to predict abundance based on the number of fish caught on the first pass and any other covariates.

Abundance estimates at all sites were translated into areal fish densities (parr/m2) 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 218 unique sites with paired fish-habitat data (Table 1).

## 2.3 Habitat Covariate Selection

A crucial step in developing the Chinook parr QRF model to predict habitat fish capacity 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 CHaMP 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). Finally, including too many habitat covariates can result in overfitting of the model (e.g., including as many covariates as data points).

To prevent overfitting the model, we paired down the more than 100 metrics generated by the CHaMP protocol. We considered first the association between the habitat metric and observed fish densities, and second the correlation among habitat metrics. We used the Maximal Information-Based Nonparametric Exploration (MINE) class of statistics (Reshef et al. 2011) to determine those habitat characteristics most highly associated with observed parr densities. We calculated the maximal information coefficient (MIC) using the R package *minerva* (Albanese et al. 2013) 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 the response variable, Chinook parr density (fish/m2), was used to inform decisions on which habitat covariates to include in the model. Habitat metrics were grouped into broad categories that include channel unit, complexity, cover, disturbance, riparian, size, substrate, temperature, water quality, and woody debris. Within each category, metrics were ranked according to their MIC value (Figure 1). Our strategy was to select one or two variables with the highest MIC score within each category so that covariates describe different aspects of the rearing habitat. Additionally, we attempted to avoid covariates that were highly correlated (Figure 2) 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 (Table 2). 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. We fit the QRF models using the *quantregForest* function from the *quantregForest* package (Meinshausen 2016) in R software (R Core Team 2015). 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 aimed at the 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 used to predict Chinook salmon parr capacity using measurements of the habitat covariates used to fit the model. In our case, this includes all sites in the Columbia River Basin surveyed using the CHaMP protocol (CHaMP 2016). 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 Buehlmann 2012; Stekhoven 2013). For CHaMP sites surveyed in multiple years, we first calculated the mean among years prior to making predictions. In total, we generated 485 predictions of spring/summer Chinook parr carrying capacity, for summer months, for xxx CHaMP sites in the Columbia River Basin. Carrying capacity predictions were made for sites in the following basins: Entiat, Grande Ronde (upper, including Minam), John Day, Lemhi, Methow, Secesh, Tucannon, Yankee Fork, and Wenatchee.

## 2.5 Model Extrapolation

Using the QRF model, we predicted habitat capacity for juvenile rearing at all CHaMP sites within the interior Columbia River basin (CRB). To predict larger scale capacity (e.g., watershed, population), we developed two extrapolation models (linear and areal) based on globally available attributes (GAA) which often corresponded to the entirety of tributary habitat utilized by a given population. The GAA data were available from a list of Generalized Random Tessellation Stratified (Stevens and Olsen 1999, 2004) master sample sites that CHaMP sample sites were initially selected from. The extrapolation model used a multiple linear regression model that incorporated the design weights of CHaMP sites using the *svyglm* function from the *survey* package (Lumley 2004, 2016) in R software (R Core Team 2015). Models with all possible combination of GAAs were fit and all models with cumulative Akaike weights less than 0.95 were model averaged to make predictions of capacity at all master sample sites throughout the interior CRB; generally spaced approximately one kilometer apart. For sites outside of CHaMP watersheds, only extrapolation models that excluded watershed as a covariate were model averaged to make predictions (Table 4).

The natural log of the CHaMP site capacity predictions were used as the response variable for the extrapolation model. We determined mean linear capacity (fish/m) of the master sample points within a given spatial scale to summarize capacity at larger scales. For visualization, predictions of areal capacity (fish/m2) were made. We only used master sample points within the domain of spring/summer Chinook salmon. The domain for a given watershed was either determined by StreamNet <http://www.streamnet.org> or using expert opinion from local biologists. The mean of capacity estimates within that scale were multiplied by the total length of the stream within the domain to estimate total capacity. For each response (linear and areal fish density), two models were developed: one for sites within CHaMP watersheds, and one for sites outside of CHaMP watersheds (Table 4).

Here, we show capacity extrapolation results for the Chinook domain within the Lemhi River watershed, a tributary of the Salmon River, Idaho, including 10 Lemhi River tributaries (Table X). Moreover, we provide maps to visualize predictions of Chinook salmon parr capacity both at CHaMP sites using the QRF model and at all master sample points within the Chinook salmon domain using the extrapolation model (Figure X).

## 2.6 Model Validation

Estimates of total Chinook salmon parr capacity were made for xx watersheds or populations throughout the interior Columbia River Basin. Observed fish data from Chinook salmon populations within the interior Columbia River that were not used to construct the QRF models were used to determine whether extrapolated capacity estimates were legitimate. Spawner-recruit data from several watersheds were compiled to validate the extrapolated QRF estimates of Chinook salmon parr capacity. Some watersheds had direct estimates of parr, while others had estimates of smolts from rotary screw traps. For watersheds only with smolt estimates, estimates of parr were calculated using over-winter survival estimates from those watersheds (Table 5). A series of spawner-recruit functions were then fit to this data including Beverton-Holt, Ricker, and hockey stick. Each of these functions also provide an estimate of parr capacity; parr capacity estimates from each of the spawner-recruit functions were compared with QRF estimates of Chinook salmon parr capacity (Table 5, Figure 4).

# 3. Results

## 3.1 Habitat Covariate Selection

We categorized 82 habitat measurements collected using the CHaMP habitat protocol (CHaMP 2016) into 10 habitat groups. Within the initial dataset containing paired fish-habitat data for 189 unique CHaMP sites (Table 1), a maximal information coefficient (MIC) value was calculated for each habitat metric based on the strength of association with observed Chinook salmon parr density (fish/m2). Covariates were then ranked within each habitat group (Figure 2) and one or two covariates within each habitat group were selected, taking into consideration MIC rank. Our strategy was to 1) consider pairwise correlations among habitat covariates (Figure 3) to minimize redundant covariates measuring similar aspects of habitat, and 2) select covariates that describe habitat characteristics likely important towards rearing of Chinook salmon parr. Considering the above, we chose the following 14 CHaMP habitat covariates to fit the QRF model: channel unit frequency, conductivity, cumulative drainage area, disturbance index, fish cover (total), large wood frequency (wetted), riparian cover (no canopy), slow water percentage, substrate <2 mm, substrate D50, summer hourly average temperature, wetted depth standard deviation, and wetted width to depth ratio average (Table 2).

## 3.2 QRF Model

A QRF model was fit using 14 habitat covariates (Table 2) and the *quantregForest* package (Meinshausen 2016) in R (R Core Team 2015). After removing records with more than three missing covariates, 186 records remained resulting in 13.3 data points per covariate. Figure 3 shows the relative importance of each habitat covariate included in the model, after model fit. Cumulative drainage area at a given site was ranked as the highest relative importance for predicting Chinook salmon parr density, whereas the frequency of large woody debris was ranked as least important among the 14 habitat covariates within the model. Additionally, we visualized the marginal effect of changing a single habitat covariate on the response variable while maintaining all other covariates at their mean values (Figure 5). Finally, the QRF model was used to predict Chinook salmon parr rearing capacity during summer months at all CHaMP sites within the interior Columbia River Basin..

## 3.3 Model Extrapolation

In total, 14 GAA habitat covariates were used among the four extrapolation models used to extrapolate CHaMP site carrying capacity predictions to larger scales (e.g., watershed, population) (Table 3). GAA habitat covariates included: mean annual velocity, slope, drainage area, bankfull width (modeled), stream power, bankfull width, channel type, temperature range, growing degree day, precipitation, elevation, CHaMP watershed, and three principal components describing disturbance and natural classifications (Whittier et al. 2011). Table 4 shows a summary of model fit for each of the extrapolation models to extrapolate QRF model predictions to larger scales.

We estimated areal fish capacity both 1) at CHaMP sites within the Lemhi River, Idaho from the QRF model prior to extrapolation and 2) across master sample sites within the Chinook salmon domain in the Lemhi River after extrapolating CHaMP site predictions across the landscape. Overall, the extrapolation model estimated a total Chinook salmon parr carrying capacity of 358,038 (SE = 74,837) for the Lemhi River watershed during summer months.

## 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 6, Figure 7). 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 7). 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 overlap among estimates. Correlations between parr capacity estimates from the QRF model and spawner-recruit models ranged from 0.874 (Ricker) to 0.996 (Beverton-Holt) (Figure 8). 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

* 1. Provides a tool to estimate habitat capacity
     1. Why is this important?
     2. Can be used to quantify habitat
     3. Used to direct habitat restoration prioritization, implementation, and evaluation
     4. Important that models can be life-stage specific
        1. Allows evaluation of where habitat is limiting
        2. Direct habitat restoration
  2. Models are currently available for Chinook salmon and steelhead
     1. Summer parr and redds
     2. Winter pre-smolt currently under development
     3. Could easily be expanded to other Pacific salmon species, resident salmonids, or even any fish species in wadeable streams
  3. Discussion of fish-habitat relationships from PDPs
     1. Biological expectations from QRF model
  4. Discussion of model assumptions (KEVIN)
     1. What assumptions are realistic?
     2. What assumptions need to be further addressed or evaluated?
  5. Extrapolation Model:
     1. Pros:
        1. Provides estimates at any larger scale given spatial extents
     2. Cons:
        1. Relies on messy GAA data
        2. CHaMP habitat data is not reproducible, expensive
        3. Correlations between habitat and capacity is messy
        4. Why else does it suck?
  6. The future (RICHIE)
     1. Channel unit scale QRF
     2. Improved habitat data collection
        1. Drones
        2. LiDAR
     3. Collect remote sensing habitat data everywhere
     4. Plug the Stream Habitat Assessment Protocol
        1. Use language from the protocol!
  7. Next Steps
     1. Drones and channel units
     2. Expand field work
        1. Primarily habitat data collection
  8. Concluding Thoughts

## 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 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 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 asses higher quantiles that are assumed to represent quality habitat.

## 4.3 Discussion of Fish-Habitat Relationships from PDPs

The results of the QRF parr capacity model for spring/summer Chinook salmon meet many biological expectations. Discussion of biological expectations and PDP plots…

## 4.4 Model Assumptions

Model assumptions should always be considered when interpreting results; the following assumptions were made when modeling habitat capacity using QRF methods. First, we assumed that sites with higher carrying capacity (i.e., ‘better’ habitat) contain more parr during fish surveys. Second, we assumed that carrying capacity is a function of habitat. Third, we assumed that estimates of parr densities from fish surveys are unbiased. Fourth, we assumed that at least some site surveys in our dataset are at or close to carrying capacity at the site level; however, if this fourth assumption is not met, we would obtain a conservative estimate of carrying capacity (but the framework of the model is not wrong). Moreover, we assumed that unmeasured factors other than habitat may be limiting fish densities from nearing carrying capacity (e.g., insufficient adult spawners, competition, predation, etc.). Finally, we assumed that the 90th quantile of predictions at each CHaMP site is a reasonable proxy for carrying capacity. With insufficient data, it is difficult to get a well-defined estimate of higher quantiles because substantial data are necessary to define the tails of a distribution.

Assumptions we don’t need to make. Because QRF naturally incorporates non-linear relationships between predictor and response variables, we make no assumptions regarding the shape of fish-habitat relationships. Rather, fish-habitat relationships are estimated from the data with no constraints Moreover, we make no assumptions regarding correlations among habitat covariates sicne the random forest framework accounts for possible interactions between habitat metrics.

## 4.5 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 estimates. 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.

## 4.6 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 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 2017).

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.

## 4.7 Next Steps

Text…

## 4.8 Concluding Thoughts

Text…

# 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.

# References

Albanese, D., M. Filosi, R. Visintainer, G. Jurman, and C. Furlanello. 2012. Minerva and minepy: a C engine for the MINE suite and its R, Python, and MATLAB wrappers. Bioinformatics. 29(3):407-408.

Breiman, L. 2001. Random forests. Machine Learning. 45(1):5-32.

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. 34:621-630.

CHaMP (Columbia Habitat Monitoring Program). 2016. Scientific protocol for salmonid habitat surveys within the Columbia Habitat Monitoring Program. Prepared by CHaMP for the Bonneville Power Administration.

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

Good, T.P., R.S. Waples, and P. Adams (editors). 2005. Updated status of federally listed ESUs of West Coast salmon and steelhead. U.S. Dept. Commer., NOAA Tech. Memo. NMFS-NWFSC-66, 598 p.

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.

ISEMP/CHaMP. 2017. Integrated Status and Effectiveness Monitoring Program (ISEMP) and Columbia Habitat Monitoring Program (CHaMP) Annual Combined Technical Report, January – December 2016, BPA Projects 2003-017-00 and 2011-006-00, 03 Electronic Pages.

Knudby, A., A. Brenning, and E. LeDrew. 2010. New approaches to modeling fish-habitat relationships. Ecological Modeling. 221(3):503-511.

Lecun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature. 521: 436-44

Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R news 2:18-22.

Lumley, T. 2016. survey: analysis of complex survey samples. R package version 3.31-5.

Lumley, T. 2004. Analysis of complex survey samples. Journal of Statistical Software. 9(1):1-19.

Meinshausen, N. 2016. quantregForest: Quantile Regression Forests. R package version 1.3-5. <https://CRAN.R-project.org/package=quantregForest>

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

Ogle, D.H. 2017. FSA: Fisheries Stock Analysis: R package version 0.8.17.

Prasad, A., L. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystem. 9(2):181-199.

R Core Team. 2015. 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. Mitzenbacher, and P.C. Sabeti. 2011. Detecting Novel Associations in Large Data Sets. Science. 334:1518-1524.

Rivest, L.P. and S. Baillargeon. 2014. Rcapture: Loglinear models for capture-recapture experiments. R package version 1.4-2.

Robson, D. and H. Regier. 1964. Sample size in Peterson 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.

Stekhoven, D.J. 2013. missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4.

Stekhoven, D.J. and P. Buehlmann. 2012. MissForest – non-parametric missing value imputation for mixed-type data. Bioinformatics. 28(1):112-118.

Stevens, D.L., Jr. and A.R. Olsen. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association. 99(465):262-278.

Stevens, D.L., Jr. and A.R. Olsen. 1999. Spatially restricted surveys over time for aquatic resources. Journal of Agricultural, Biological, and Environmental Statistics. 4:415-428.

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(1):104-117.

Weinstein, B.G. (2017). A computer vision for animal ecology. Journal of Animal Ecology.87: 533-545

Whitehead, K. Hugenholtz, C.H. 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.

Whittier, T., A. Herlihy, C. Jordan, and C. Volk. 2011. Landscape classification of Pacific Northwest hydrologic units based on natural features and human disturbance to support salmonid research and management. NOAA, National Marine Fisheries Service. NOAA Contract #AB1133F10SE2464. Pp. 39.

# Tables

Table . The number of unique sites in the initial dataset, by CHaMP watershed, with paired fish-habitat data used to populate the Chinook salmon parr capacity model.

|  |  |
| --- | --- |
| **CHaMP Basin** | **n** |
| Entiat | 36 |
| Grande Ronde (upper) | 84 |
| John Day | 26 |
| Lemhi | 18 |
|  |  |
| Minam | 9 |
|  |  |
|  |  |
| South Fork Salmon | 26 |
| Wenatchee | 19 |
| **Total:** | **189** |

Table . Habitat metrics and descriptions of metrics included in the QRF model to predict spring/summer Chinook salmon parr capacity (during summer months). Metrics are ranked in order of relative importance.

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Metric | Metric Category | Description |
| 1 | Cumulative Drainage Area | Size | The cumulative land area that drains to a given location/site. |
| 2 | Wetted Width Integrated | Size | Average width of the wetted polygon for a site. |
| 3 | Wetted Depth SD | Complexity | Standard Deviation of water depths within the wetted channel. |
| 4 | Conductivity | Water Quality | Measure of concentration of ionized materials in water, or the ability of water to conduct electrical current. |
| 5 | Summer Hourly Average Temp | Temperature | Average of all hourly temperature measurements collected July 15th - August 31st. |
| 6 | Wetted Width to Depth Ratio Avg | Complexity | Average width to depth ratio of the wetted channel measured from cross-sections. Depths represent an average of depths along each cross-section. |
| 7 | Fish Cover: Total | Cover | Percent of wetted area with the following types of cover: aquatic vegetation, artificial, woody debris, and terrestrial vegetation. |
| 8 | Channel Unit Frequency | Channel Unit | Total number of channel units per 100 meters. |
| 9 | Slow Water Percent | Channel Unit | Number of Slow Water/Pool channel units per 100 meters. |
| 10 | Riparian Cover: No Canopy | Riparian | Percent of riparian canopy devoid of vegetation. |
| 11 | Substrate: D50 | Substrate | Diameter of the 50th percentile particle derived from pebble counts. |
| 12 | Disturbance Index | Disturbance | Disturbance index that includes measures of % urban, % agricultural, % impervious surface, and road density (Whittier et al. 2011). |
| 13 | Substrate: <2 mm | Substrate | Average percentage of pool tail substrates comprised of fine sediment <2 mm. |
| 14 | Large Wood Frequency: Wetted | Wood | Number of large wood pieces per 100 meters within the wetted channel. |

Table . 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). An ‘X’ denotes the GAA was used in the given extrapolation model. Linear fish densities (fish/m) were used to extrapolate estimates to larger scales (e.g., watershed, population) whereas areal fish densities (fish/m2) were used for visualizations.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Covariate** | **Scale** | **Units** | **CHaMP (per m)** | **non-CHaMP (per m)** | **CHaMP (per m2)** | **non-CHaMP (per m2)** |
| Mean Annual Velocity | Reach (2 km) | fps | - | X | X | X |
| Slope | Reach (2 km) | m/m | - | - | X | X |
| Drainage Area (sqrt) | Reach (2 km) | km2 | - | X | X | - |
| Bankfull Width – modeled | Site (300 m) | m | X | X | - | - |
| Stream Power | Reach (2 km) | - | - | - | - | - |
| Bankfull Width | Site (300 m) | m | X | - | X | X |
| Channel Type | Site (300 m) | - | X | X | - | X |
| Temperature Range | Reach (2 km) | °C | - | - | - | X |
| Growing Degree Day | Reach (2 km) | GDU | - | X | X | X |
| Precipitation | Reach (2 km) | cm | - | - | - | X |
| Elevation | Site (300 m) | m | X | X | X | X |
| CHaMP Watershed | Region | - | X | - | X | - |
| Disturbance Class PCA 1 | Watershed (HUC12) | - | X | X | - | X |
| Natural Class PCA 1 | Watershed (HUC12) | - | - | X | X | X |
| Natural Class PCA 2 | Watershed (HUC12) | - | X | - | - | - |

Table . Summary of model fit for each of the extrapolation models to extrapolate quantile regression forest (QRF) model predictions of spring/summer Chinook parr capacity to larger scales (e.g., watershed, population).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Response** | **R2** | **Adjusted R2** | |
| CHaMP | fish/m | 0.493 | 0.466 |
| non-CHaMP | fish/m | 0.398 | 0.374 |
| CHaMP | fish/m2 | 0.458 | 0.434 |
| non-CHaMP | fish/m2 | 0.407 | 0.382 |

Table . Estimates of total Chinook salmon parr capacity, during summer months, for the Lemhi River, Idaho and tributaries with standard error.

|  |  |  |  |
| --- | --- | --- | --- |
| Stream | Stream Length (km) | Parr Capacity | SE |
| Big Springs Creek | 6.8 | 15,159 | 4,293 |
| Big Timber Creek | 2.0 | 11,102 | 0 |
| Bohannon Creek | 1.4 | 3,950 | 892 |
| Canyon Creek | 6.2 | 17,946 | 4,281 |
| Hayden Creek | 19.9 | 24,123 | 11,789 |
| Kenney Creek | 1.5 | 3,176 | 0 |
| Lemhi River | 99.8 | 265,739 | 47,828 |
| Little Springs Creek | 5.0 | 13,323 | 2,209 |
| Wimpey Creek | 2.8 | 3,520 | 1,282 |
| Total: | 145.4 | 358,038 | 74,837 |

Table . 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 blank. ODFW = Oregon Department of Fish and Wildlife; WDFW = Washington Department of Fish and Wildlife; IDFG = Idaho Department of Fish and Game; RST = Rotary Screw Trap; QCI = Quantitative Consultants, Inc.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Watershed** | **n Years** | **Adult Data** | **Adult Source** | **Parr Data** | **Parr Source** | **Beverton-Holt** | **Ricker** | **Hockey Stick** | **QRF** |
| Catherine Creek | 20 | Spawners | ODFW | RST | ODFW | 135,387 (0.269) | 103,021 (0.141) | 99,921 (0.210) | 179,004 (0.039) |
| Chiwawa River | 20 | Spawners | WDFW | Parr Surveys | BioAnalysts | 248,586 (0.240) | 166,139 (0.148) | 174,216 (0.184) | 341,352 (0.059) |
| Grand Ronde River (upper mainstem) | 8 | Spawners | ODFW | RST | ODFW | 171,609 (0.388) | 168,137 (0.298) | 127,052 (0.317) | 212,794 (0.045) |
| Hayden Creek | 8 | Redds | IDFG | RST | IDFG/QCI | 58,493 (0.726) | 37,235 (0.371) | 41,668 (0.535) | 57,651 (0.063) |
| Minam River | 14 | Spawners | NA | RST | NA | NA (NA) | 484,810 (1.444) | 662,806 (1.726) | 262,348 (0.049) |
| Tucannon River | 27 | Redds | WDFW | RST | WDFW | NA (NA) | NA (NA) | 152,809 (0.165) | 380,301 (0.029) |
| upper Lemhi River | 22 | Redds | IDFG | RST | IDFG/QCI | 112,746 (0.784) | 64,303 (0.503) | 72,470 (0.638) | 150,706 (0.061) |

# Figure Captions

Figure . Map of the Columbia River Basin showing each of the Columbia Habitat Monitoring Program (CHaMP) watersheds.

Figure . Barplot showing the strength of association between each habitat metric and observed densities (fish/m2) of spring/summer Chinook salmon parr, during the summer, facetted by habitat categories. Results were used to determine which the habitat covariates to include in the quantile regression forest (QRF) model to predict carrying capacity.

Figure . Pairwise correlations among habitat covariates used to fit the spring/summer Chinook salmon parr capacity quantile regression forest (QRF) model.

Figure . 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 . 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, and the sub-basin they came from.

Figure . Predictions of Chinook salmon parr carrying capacity (fish/m2), during summer months, at CHaMP sites within the Lemhi River, Idaho based on 14 habitat covariates and a quantile regression forest (QRF) model (left). Estimates at CHaMP sites were then extrapolated across all master sample sites in the Lemhi River using an extrapolation model and globally available attribute (GAA) data (right).

Figure . 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.

Figure . Scatterplots showing correlations among estimates of Chinook salmon parr carrying capacity from the quantile regression forest (QRF) model and three spawner-recruit models.

# Figures

Figure 1

Figure 2

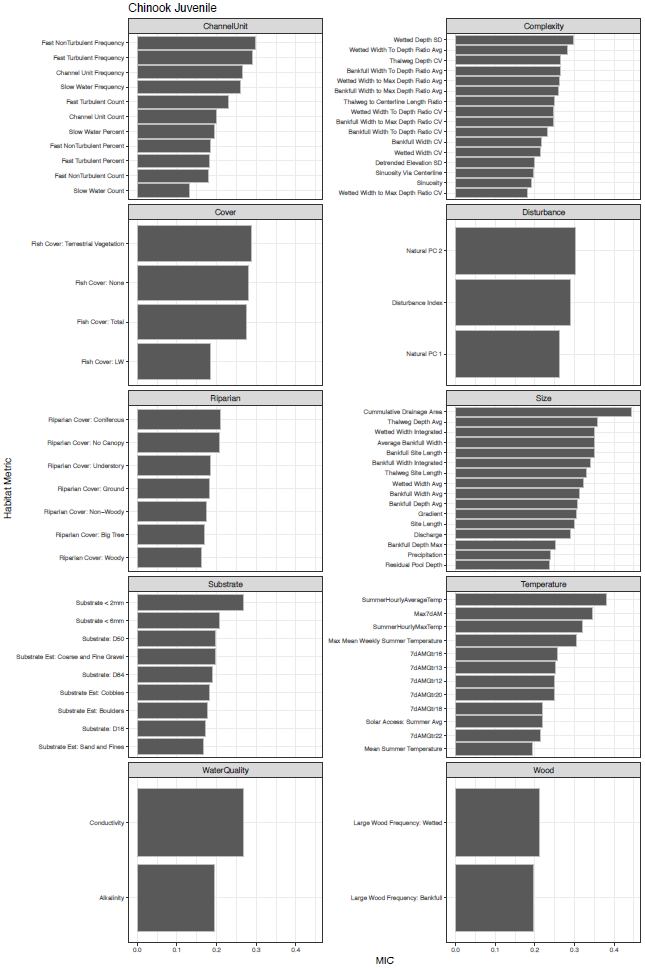


Figure 3

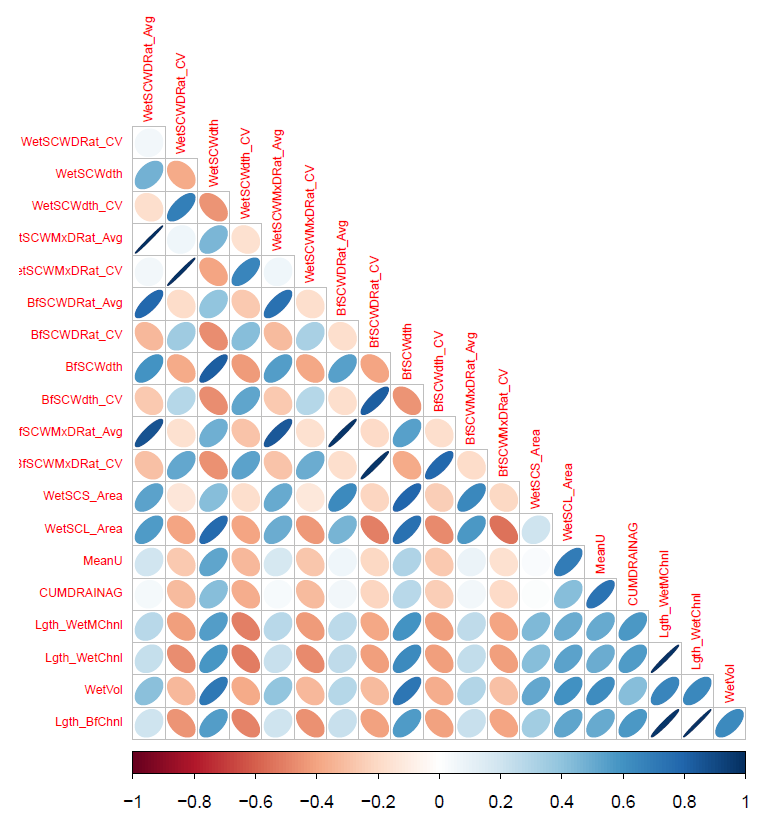


Figure 4

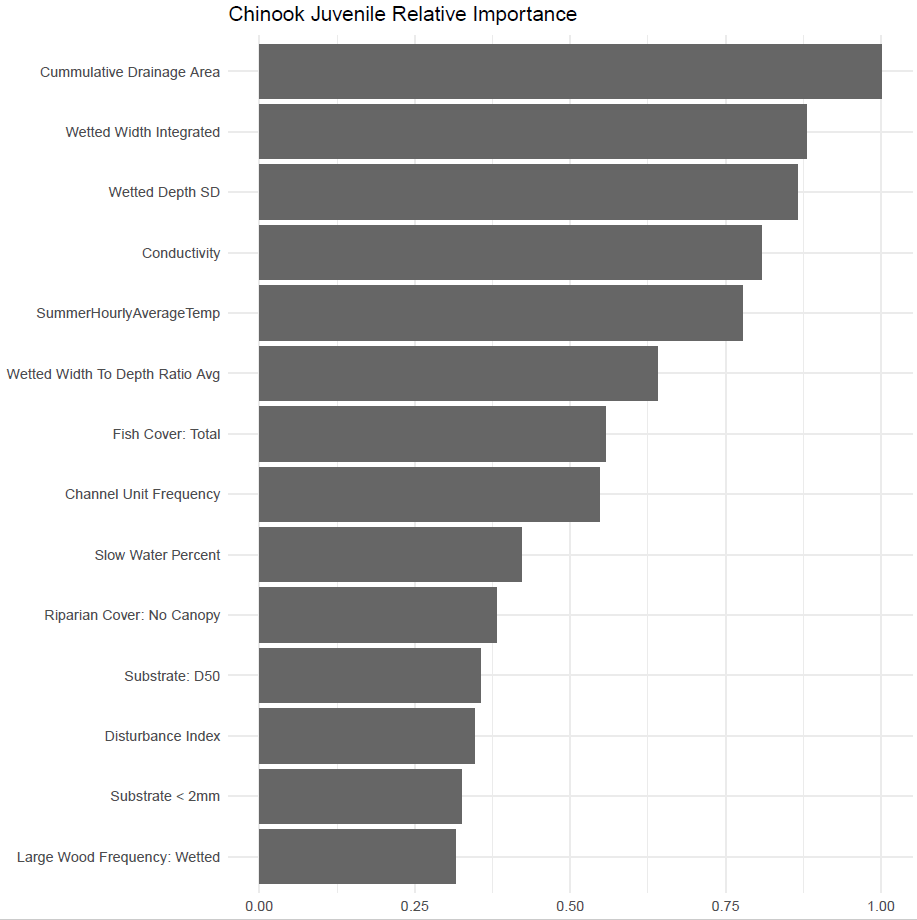


Figure 5

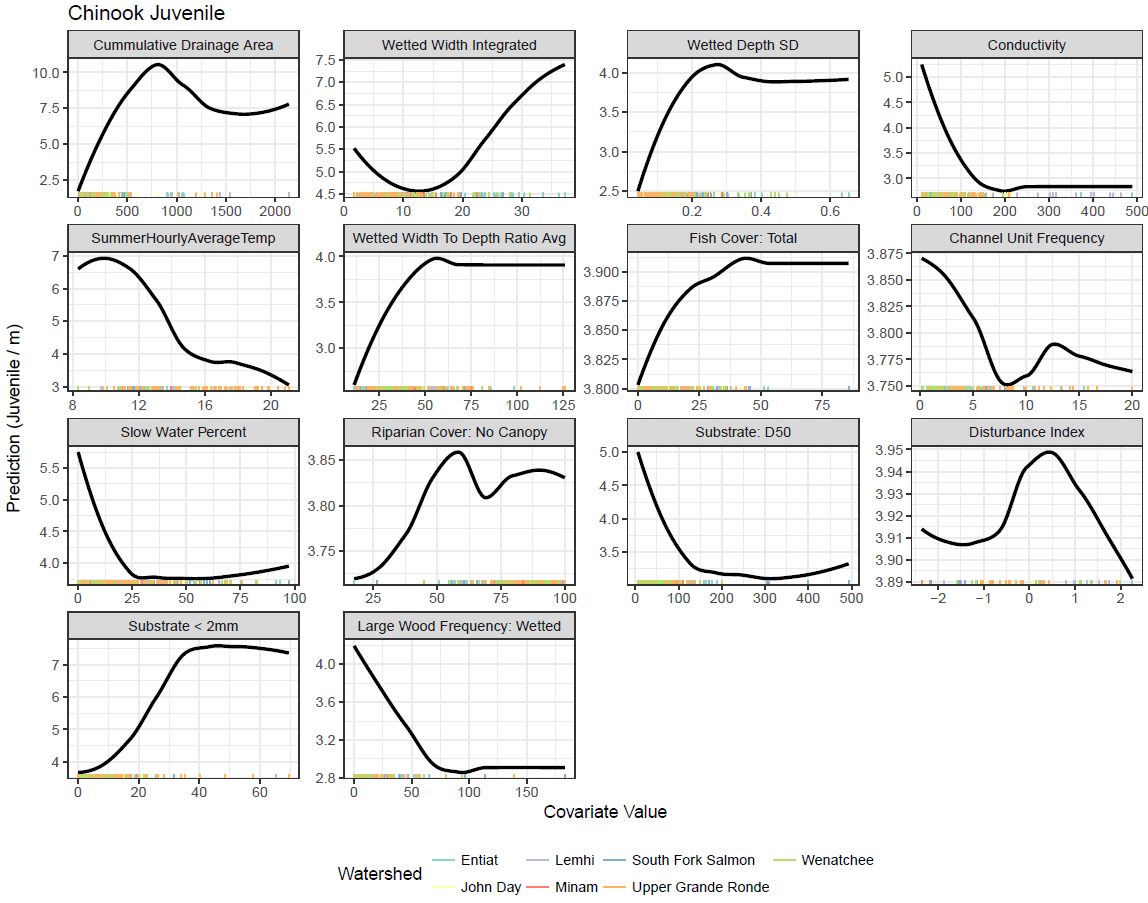


Figure 6

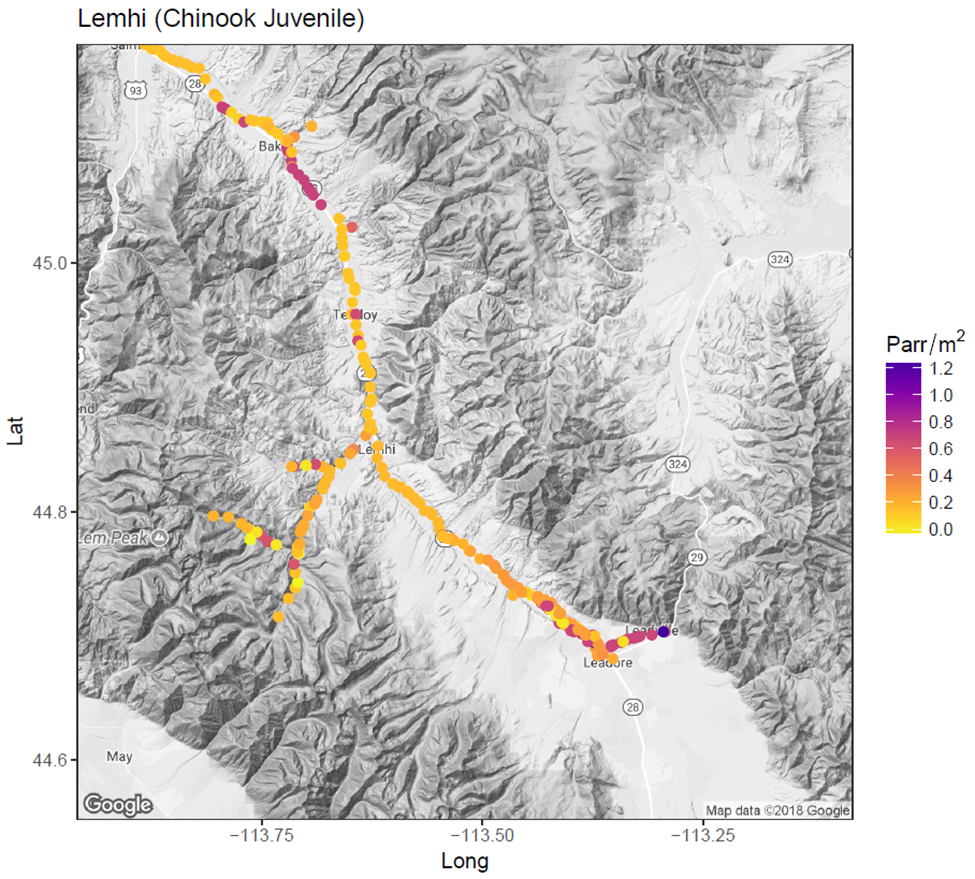
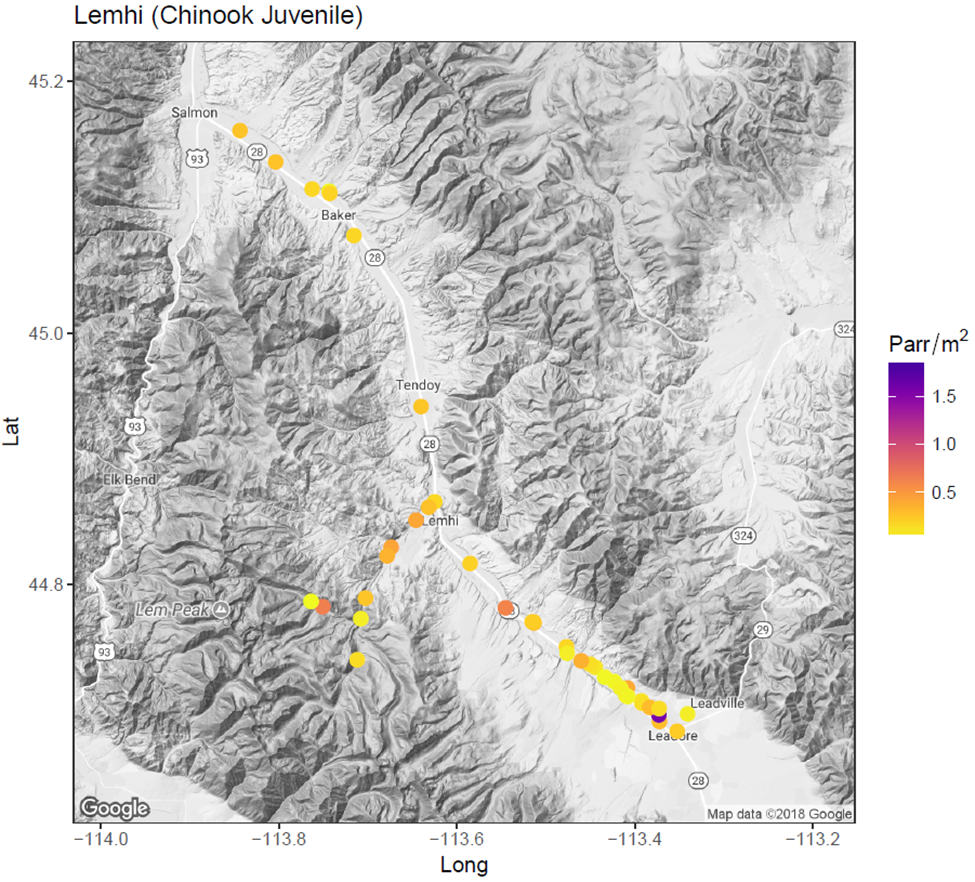
****

Figure 7

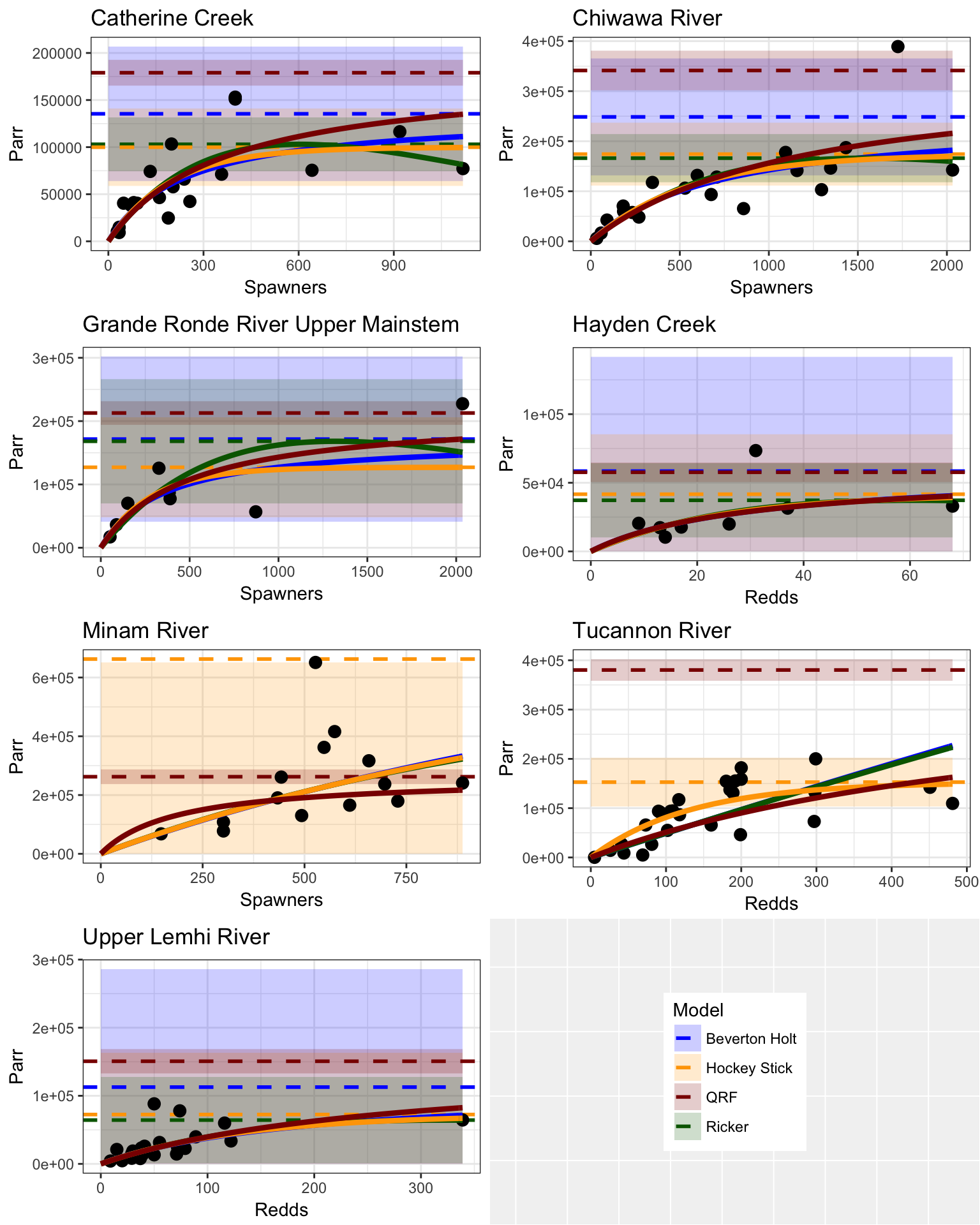


Figure 8

