SHORT COMMUNICATION

HABITAT SUITABILITY CRITERIA VIA PARAMETRIC DISTRIBUTIONS: ESTIMATION, MODEL SELECTION AND UNCERTAINTY

NICHOLAS A. SOM[a,d]*, DAMON H. GOODMAN[a], RUSSELL W. PERRY[b] AND THOMAS B. HARDY[c]

[a] *Arcata FWO, US Fish & Wildlife Service, Arcata, California, USA*
[b] *Western Fisheries Research Center, US Geological Survey, Cook, Washington, USA*
[c] *Meadows Center for Water and the Environment, Texas State University, San Marcos, Texas, USA*
[d] *Department of Fisheries Biology, Humboldt State University, Arcata, California, USA*

ABSTRACT

Previous methods for constructing univariate habitat suitability criteria (HSC) curves have ranged from professional judgement to kernel-smoothed density functions or combinations thereof. We present a new method of generating HSC curves that applies probability density functions as the mathematical representation of the curves. Compared with previous approaches, benefits of our method include (1) estimation of probability density function parameters directly from raw data, (2) quantitative methods for selecting among several candidate probability density functions, and (3) concise methods for expressing estimation uncertainty in the HSC curves. We demonstrate our method with a thorough example using data collected on the depth of water used by juvenile Chinook salmon (*Oncorhynchus tschawytscha*) in the Klamath River of northern California and southern Oregon. All R code needed to implement our example is provided in the appendix. Published 2015. This article is a U.S. Government work and is in the public domain in the USA.

INTRODUCTION

Coupled physical and biological models often focus on interactions of target species with their environment for the purposes of exploring emergent properties (Harvey and Railsback, 2009), evaluating management scenarios (Sandoval-Solis *et al.*, 2013) or predicting effects of future climate change (Holsinger *et al.*, 2014). Key inputs to these models are the information on the association of species with characteristics of their habitats. By quantifying species' habitat characteristics, these dynamic models simulate population responses to physical habitat change. As such, an essential component of model construction is mathematically linking the gradient in habitat quality to the model's spatial domain as a function of physical variables.

Many methods exist to model habitat quality as a function of physical attributes (see review by Ahmadi-Nedushan *et al.*, 2006). These often result in a habitat suitability index (HSI) that ranges between zero and one. Values closer to one indicate desirable habitats likely to be selected by the focal species, while values closer to zero suggest poor quality habitat that may be avoided. The suite of available methods cover presence–absence data (Beakes *et al.*, 2014) that represent locations where the species is both present and absent or presence-only data (Elith *et al.*, 2011; Royle *et al.*, 2012) with measurements only at locations of species occupation. In addition to these differing response data structures, modellers may consider a multivariate approach, where the effects of physical variables are modelled simultaneously (Ahmadi-Nedushan *et al.*, 2006), or a univariate approach, where the effect of each variable is first analyzed independently (Hayes and Jowett, 1994), and then a suitable method (e.g. geometric or arithmetic mean (Hardy *et al.*, 2006a), multiplication (Gard, 2014), etc.) is used to combine the univariate analyses into a single HSI value.

The univariate approach often leads to habitat suitability criteria (HSC) curves, which express the relative quality of a variable across its range. HSC values also lie between zero and one and are interpreted similarly to HSI. For example, in the case of riverine fish, it is not uncommon to relate fish use with the depth and velocity of water, and the distance to nearest cover (Persinger *et al.*, 2011). If habitat availability (relative amounts of the physical variable across the entire sampling domain) information is available, then HSC curves

*Correspondence to: Nicholas A. Som, US Fish & Wildlife Service, Arcata FWO, Arcata, CA, USA.
E-mail: nicholas_som@fws.gov

can be generated by adjusting habitat use by the availability of each physical variable (Jowett, 2002), via methods that include forage ratios (Heath *et al.*, 2015).

Univariate HSC curves have been generated using methods ranging from professional judgement to quantitative analysis (Bovee, 1986; Newcomb *et al.*, 2007) and among the most commonly applied and cited methods for HSC curve creation is that of Hayes and Jowett (1994; e.g. cited over 100 times at the time of submission). This method relies on kernel-smoothed density distributions, which use both a kernel smoother (i.e. the shape of smoothing window) and a smoothing parameter (also called bandwidth) to approximate the distribution of the observed data. The smoothing parameter is selected to match the pattern of the observed data without masking important components or shapes and can be aided by variable smoothers that adapt the amount of smoothing for long-tailed distributions (Silverman, 1986). Hence, applying kernel smoothing requires the user to select both a smoothing parameter and a smoothing shape (there are many), but this method lacks quantitative methods to evaluate the choices.

There are several applications of this method, with varying steps after the kernel-smoothed density curves are created. For instance, the HSC curves could be adjusted for availability, standardized to lie between zero and one and then matched to mathematical functions via maximum likelihood using coordinates of the kernel-density curves (Hayes and Jowett, 1994). The mathematical functions can be polynomials of varying orders or other non-linear functions. In essence, this method first serves to approximate the probability density function of the observed data via a kernel-smoothed density distribution and then attempts to describe the resulting curve via a mathematical function that can be used to assign HSC curve values across the range of potential values in the system of study.

We agree with Hayes and Jowett (1994) that probability density functions (hereafter, PDFs) are an elegant representation of HSC curves. We contend, however, that given the large array of potential PDFs (and large suite of potential shapes) that these data are more amenable to direct PDF fitting and estimation. Herein, we propose the direct use of PDFs as the basis for the mathematical representation of HSC curve shapes. In addition to eschewing decisions regarding kernel shape or smoothing parameter size, we show that our proposed method has the added benefits of (1) estimating parameters directly from the observed data (instead of fitting a mathematical function to kernel-density output), (2) providing a clear method for quantitative evaluation of candidate PDFs, and (3) creating a framework where expressing the uncertainty in parameter estimates and PDF choice is easily implemented with common statistical software.

PROBABILITY DENSITY FUNCTIONS FIT TO HSC CURVES

Continuous probability density functions

We begin with a very concise introduction to continuous PDFs. Interested readers can find more details in any mathematical statistics text, including Mukhopadhyay (2000). In general, for any continuous random variable X, $f(x)$ is a probability density function if and only if $f(x) \geq 0$ for all real x, and $\int_{-\infty}^{\infty} f(x)dx = 1$. These conditions are assured if applying common software for fitting distributions to data, such as R statistical software (R Core Team, 2014). There are many continuous PDFs, and common distributions include the normal (also called Gaussian), gamma, chi-square, exponential, Pareto and Weibull. For variables like depth, velocity, distance to cover and other physical variables that might be combined into an HSC analysis, the variables tend to be non-negative, and other potential continuous PDFs include the kappa, generalized extreme value, log-normal and truncated (at zero) normal distribution.

The support (i.e. the range of all possible values that X can take), the range of values X is most likely to take and the shapes of PDFs are governed by the form of $f(x)$ and the parameters that define the function. For example, the normal distribution has support from $-\infty$ to ∞, and the gamma is only defined for $x \geq 0$, but shapes and likely ranges of X from PDFs can vary demonstrably with differing parameter values (Figure 1). In all, this
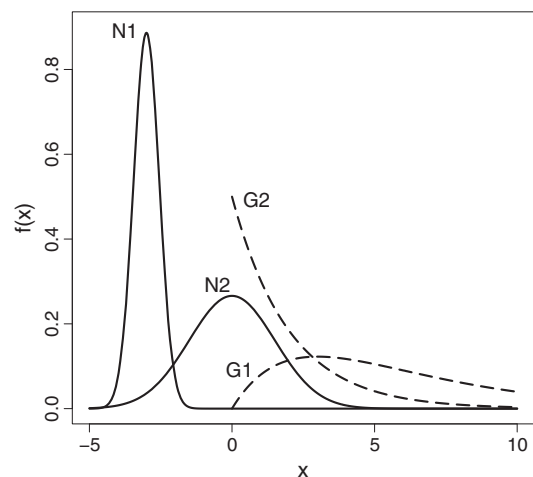
Figure 1. Example probability density functions of two normally distributed random variables (N1 and N2, solid lines) and two gamma random variables (G1 and G2, dashed lines). N1 has parameter values mean = −3 and standard deviation = 0.45; N2 has parameter values mean = 0 and standard deviation = 1.5; G1 has parameter values shape = 1 and scale = 2; G2 has parameter values shape = 2 and scale = 3. Definitions of these probability density functions and their parameters can be found in Asquith (2014)

leads to a rich library of potential PDFs from which one or several are likely to fit observed data well.

PDF parameter estimation

To take advantage of the shapes of PDFs for HSC curves, PDF parameters must be estimated from the observed data. There are several methods for estimating PDF parameters. Maximum likelihood (ML) estimation is likely the most familiar to readers, but L-moments estimation (based on linear combinations of order statistics) may be preferred for some PDFs (Hosking, 1990). Here, we present a small example of ML estimation, which finds parameter values that are the most likely (have the maximum likelihood) given the observed data, and our expanded example in the subsequent discussion demonstrates an application of L-moments estimation.

Using calculus and algebra, parameters of some PDFs may be estimated via ML analytically using summary statistics of the observed data, but there are often situations where numerical methods must be employed. Numerical methods are quite accessible with common statistical software, so we present our example using R statistical software (R Core Team, 2014), with data randomly generated from an exponential distribution. More details on ML estimation can be found in many statistical texts, including Mukhopadhyay (2000).

A likelihood function ($L(\theta)$) is a function of the parameters (θ) and is simply the product the PDF ($f(x)$) evaluated at the n data points for θ,

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

For computational convenience, the natural logarithm (*ln*) of the joint distribution of the data is often used, which results in a summation over the joint distribution. Additionally, because the optimization routine employed in our example (*optim*) defaults to finding minimums, and for a PDF selection metric described subsequently, we also multiply our function by −1 and obtain a function to use for computing ML estimates of the PDF parameters,

$$-lnL(\theta) = -\sum_{i}^{n} ln(f(x_i; \theta)) \qquad (1)$$

In the R code below this paragraph, we (1) randomly generate 45 exponentially distributed variables (rate parameter = 5); (2) place the PDF (*dexp* for the base package of R) and parameter in Equation (1); and (3) provide a reasonable parameter starting value and have the software numerically obtain a parameter estimate.

```
> #1
> set.seed(34593); data <- rexp(45, 5)
> #2
> expML <- function(pars, data)
  {-sum(log(dexp(data, rate = pars[1])))}
> #3
> start <- mean(data); fitEXP <- optim(start,
  expML, data=data, method="BFGS")
> fitEXP$par
[1] 5.172151
```

The estimate of around 5.17 is not exactly equal to 5, and that is because a random sample was drawn. If we repeated this exercise many times, the average estimate would be very close to 5. R statistical software contains many packages for estimating the parameters of PDFs via ML (e.g. *maxLik, likelihood and bbmle*) or L-moments (*lmomco*). The values of the selected PDF across the range of the variable values being considered are not restricted to be between zero and one. To impose this restriction, one simply scales the curve's values by the maximum PDF value across the range of variable values. This approach is demonstrated subsequently.

Selecting PDFs

Several PDFs will likely have potential to sufficiently fit the shape of the data, and an important step is deciding which to consider. A benefit of our method is the ability to quantitatively compare the fit of several candidate PDFs, and selecting this candidate list begins with choosing PDFs with support, and potential shape, that are appropriate for the data. For instance, a normal distribution would not be appropriate for a variable that can only take positive values, and choosing a single-parameter exponential distribution would not be appropriate for data that exhibit HSC values near zero for both the lowest and highest values of the data range. We note that our approach will work for any PDF, including those for bivariate distributions.

With our proposed method, the fit of each candidate PDF to the data can be evaluated and used to rank their relative performance. One common approach would be to compute Akaike's Information Criterion (AIC) and rank the PDFs by their AIC values (Burnham and Anderson, 2004). AIC adds a penalty for each estimated parameter and seeks to balance the number of estimated parameters against the quality of the fit. In our experience, we have not been worried about an unwieldy number of estimated parameters for these PDFs (often between one and four parameters), so we simply extract the part of AIC that does not include a penalty for the number of parameters: $-2 * ln(L(\theta))$. As two is a constant scaler in this application, we can further

simplify our metric to $-ln(L(\theta))$ (hereafter, $-LL$) as given in Equation (2). Similar to applications of AIC, the PDFs with the smallest $-LL$ would demonstrate evidence of best fit. $-LL$ is easily obtained from any of the packages for ML estimation listed earlier and also from our example with the exponential distribution. Because we incorporated both the negative and ln into our function, we simply call the function that was used for estimation and insert the estimated value.

```
> expML(fitEXP$par, data)
[1] -28.94799
```

Computing $-LL$ in this way could also be performed to compare the fits of several PDFs if another estimation method was used (e.g. L-moments), because we are simply computing a value that considers the fit of the PDF to the data given specific parameter values.

Expressing uncertainty in HSC curves

Another benefit of our method is the ability to express, both statistically and graphically, the uncertainty associated with the parameter estimates and the resulting HSC curve. Although confidence intervals for the parameter estimates are easily obtained, of more interest to practitioners is how this uncertainty is propagated to the HSC curves. The sample size and quality of fit will dictate how variant a confidence interval envelope appears around the curve generated from the observed data.

To compute and display HSC curves with their associated confidence interval curves, one could apply a parametric or non-parametric bootstrap routine. We opt here to present a non-parametric bootstrap method as it is more general, requires less knowledge of the output from various software routines and applies to a wide array of estimation techniques. Non-parametric bootstrapping involves computing statistics of interest with data generated by resampling with replacement from the observed data (Lange, 1999, p. 300). The process is repeated a large number of times (e.g. 2000) to obtain a bootstrapped approximation to the sampling distribution of the statistics of interest. Here, the statistics of interest will often be the HSC curve values across the range of the data. For an application to HSC curves, one can resample the data, estimate the parameters, generate values from the PDF over the range of the target variable, and repeat to obtain a bootstrapped distribution of HSC values that can be plotted with the curve fit to the observed data. An example of this is provided subsequently.

EXAMPLE

Here, we provide an example of our method using data collected to estimate the physical characteristics of habitat used by Chinook salmon (*Oncorhynchus tschawytscha*) fry (juvenile fish that are several months old and less than 50 mm in length) in the Klamath River of northern California and southern Oregon, USA. These are presence-only data, and we focus here on the depth variable, which is the measured depths of water occupied by these fish during data collection. These data were collected between 1998 and 2005 and contain 1663 depth values that ranged from 0.03 to 1.4 m. More details on the methods and locations of data collection, and the suite of other physical variables collected can be found in Hardy *et al*. (2006b). These data are currently being used to construct HSC curves for use in development of flow-to-habitat relationships for applications to a fish population dynamics model.

A histogram reveals a slightly right-skewed pattern, and because negative depths are not possible, only PDFs allowing only positive values are considered (Figure 2). Although we suggest that histograms be used for initial graphical assessment of the distribution of the data and for graphical assessment of how well a chosen PDF fits, we stress that we do not use histograms for the estimation of parameters or actual fitting of the HSC curves. We begin with selecting a pool of candidate PDF curves that may fit the data well, and although certainly not exhaustive, some examples include the Rayleigh, gamma and Rice distributions. Full mathematical descriptions of these PDFs can be found in Asquith (2014).

To evaluate the fit of each of these distributions to our data, we first estimate the parameters from these distributions with our data and then evaluate the $-LL$ for each fit (Table I). The R code used to complete this step is available in Appendix A.1 and relies on the R contributed package
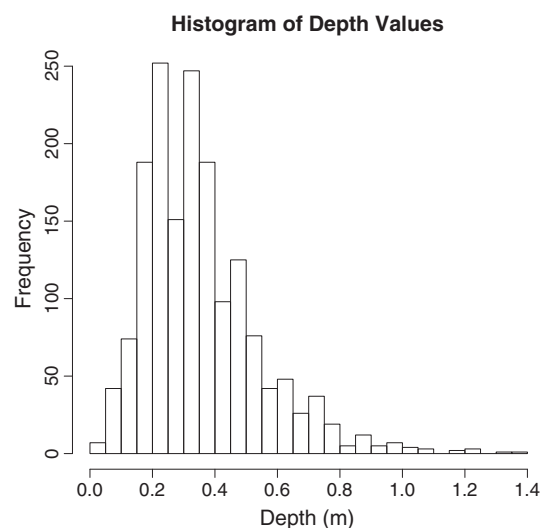

Figure 2. Histogram of measured depth values recorded with observed juvenile Chinook salmon use in the Klamath River

Table I. Summaries of parameter estimates and minus the log-likelihood (−LL) for each probability density function being considered for a habitat suitability criteria curve for the depth variable

Distribution	Parameter estimates	−LL
Gamma	$\alpha = 3.7730$; $\beta = 0.0940$	−637
Rayleigh	$\xi = 0.0144$; $\alpha = 0.2714$	−592
Rice	$v = 0.2342$; $\alpha = 0.2259$	−561

lmomco (Asquith, 2014). Based on these results, we would select the gamma PDF, because it has the lowest − LL.

Our next step is to scale the values of the selected PDF to lie between zero and one and then plot the resulting curve to visually evaluate if our estimated curve fits the data well. The R code for this step is provided in Appendix A.2. First, values for the selected PDF are computed across the observed range of depth values. Next, all PDF values are divided by their maximum to scale all values between zero and one. These scaled values are plotted against a representation of the raw data (e.g. a histogram) to ensure the curve fits well and makes sense. In our case here, the gamma distribution appears to fit the data very well (Figure 3, solid line).

Finally, we estimate how the uncertainty in our parameter estimates is reflected in the shape of the HSC curve. We apply a bootstrap routine that estimates the PDF parameters with a resampled data set and computes the HSC values over
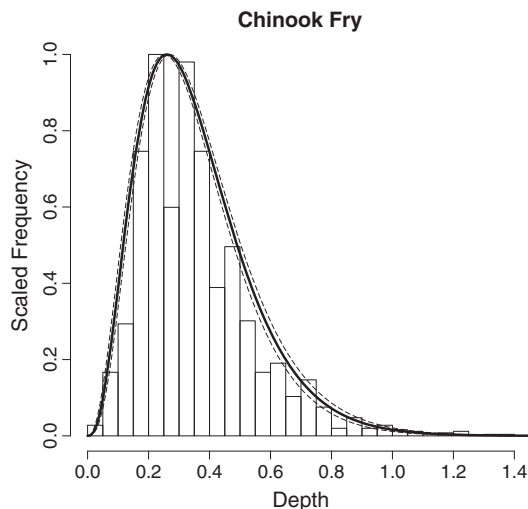

Chinook Fry

Figure 3. Estimated habitat suitability criteria (HSC) curve (solid line) with 95% confidence intervals (dashed lines) for the depth of habitats used by Chinook salmon fry in the Klamath River. The estimated HSC curve is generated by fitting the data to a gamma distribution and then scaling the fitted PDF values by their maximum to restrict values between zero and one. The 95% confidence interval curves are generated by a non-parametric bootstrapping routine. The histogram is provided for visual assessment of fit only and was not used to fit the HSC curve to the data

the range of the data. We applied this procedure 2000 times and generated an estimated sampling distribution of the HSC curve across the range of the data. Confidence intervals are generated by taking the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrapped values across the range of depth. We opted for 95% confidence intervals and, thus, took the 0.025 and 0.975 quantiles. The R code used to generate our confidence intervals is provided in Appendix A.3. In this case, the confidence intervals (Figure 3, dashed lines) track very closely, and are quite narrow around, our estimated curve. In addition to the chosen PDF fitting the data well, the confidence intervals are narrow due to the large sample size. If the sample size was much smaller, say 50, the width of the confidence intervals would be much wider and the shape less precise (e.g. Figure A.1).

DISCUSSION

We have demonstrated that PDFs can be used to construct HSC curves that are completely data-driven and without the subjectivity or reliance on smoothing parameters or histogram bin widths of other methods. The continuous PDFs employed by our method generate smooth curves, which we believe better reflect changes in habitat suitability compared with curves that vary sharply over small gradients in physical attributes, for example, frequency analysis or non-parametric tolerance limits (Newcomb *et al.*, 2007).

We are not aware of methods to adequately represent the uncertainty associated with generating HSC curves via the kernel-density approach, although it may be possible. One would need to quantitatively account for the uncertainty associated with decisions regarding type of kernel smoother employed and the amount of smoothing applied (size of bandwidth parameter). Further, the user defines the specific values, over the range of the physical variable, for which output from the kernel-smoothed density estimate is generated. This decision point precludes the use of any quantitative measures regarding quality of fit, and these values are in turn used to estimate the parameters of the mathematical function. Effectively accounting for the uncertainty in HSC curves generated using the kernel-density approach would require resolution of these issues.

We have shown that our approach provides basic methods to evaluate candidate PDFs and quantitatively select a PDF from which to base a HSC curve. Further, our approach provides straightforward methods for estimating the uncertainty in PDF parameter estimates and the resulting effects on the HSC curves. Faced with several candidate PDFs that rank similarly in regards to fit quality, one could also incorporate the uncertainty in the PDF selection to their HSC curves. For instance, at each bootstrap iteration, one could bootstrap the data as well as pull from the candidate list of

PDFs and generate HSC curves with confidence intervals reflective of both PDF and parameter uncertainty. Likewise, one could add a step to the bootstrap routine where the resampled data set is subjected to the PDF selection procedure.

If the univariate HSC curve approach is applied, there still remains the task of combining the univariate curves into a single HSI. While we do not pursue that topic here, we do note that our approach allows a straightforward way to incorporate the variation in PDF parameter estimates via common procedures like bootstrapping or Monte Carlo sampling. For instance, once could implement a routine where, at each bootstrap iteration, a bootstrapped HSC curve is created for each physical variable, and the univariate curves are combined per chosen method to create HSI values that express the according uncertainty.

We certainly do not suggest that construction of HSC curves can be made with a black-box type approach. Simply because one PDF fits best among those considered does not imply the PDF fits the data well enough for applications. In our experience, however, we have not encountered situations where a proper candidate PDF fit poorly. In fact, for applications that have covered multiple physical variables for multiple species and life stages from several rivers, we have found that gamma, kappa, Pareto and Weibull PDFs commonly fit well for HSC data. These PDFs can provide a large array of possible shapes given combinations of their respective parameter values.

We acknowledge that in our example we did not make adjustments to adjust habitat use for availability. In the habitat suitability work that we have encountered, we have not had the benefit of additional availability information. This situation could be increasingly common, with the growing need to gather information from existing databases to reduce expenses (Isaak *et al.*, 2014). Many scientists do have these data and prefer to incorporate this information into the HSC curves. Our method will work in either case. One could simply repeat the process used for selecting and estimating the parameters of a habitat-use PDF for the habitat-availability data, and then divide the use curves by the availability curves, or make availability adjustments to the use data and fit a PDF to data already adjusted for availability.

Generally, presence–absence data is preferred to presence-only, because observed zeros are informative about the suite of physical conditions that species use (Royle *et al.*, 2012). Despite this and advances in the data collection, model fitting, and analysis methods suitable for HSI creation, there will remain practitioners and applications that rely on presence-only data, by choice or by need, to construct HSC curves. We contend that the science in this arena benefits from a method that reduces subjectivity, estimates HSC curves directly from raw data and accounts for uncertainty in HSC curve creation.

REFERENCES

Ahmadi-Nedushan B, St-Hilaire A, Berube M, Robichaud E, Thiemonge N, Bobee B. 2006. A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. *River Research and Applications* 22: 503–523.

Asquith WH. 2014. lmomco|L-moments, trimmed L-moments, L-comoments, censored L-moments, and many distributions. R package version 2.1.1.

Beakes MP, Moore JW, Retford N, Brown R, Merz JE, Sogard SM. 2014. Evaluating statistical approaches to quantifying juvenile Chinook salmon habitat in a regulated California river. *River Research and Applications* 30: 180–191.

Bovee KD. 1986. Development and evaluation of habitat suitability criteria for use in the instream flow incremental methodology. Technical Report Instream Flow Information Paper 21, U.S. Fish and Wildlife Service.

Burnham KP, Anderson DR. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research* 33(2): 261–304.

Elith J, Phillips SJ, Hastie T, Dudk M, Chee YE, Yates CJ. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17(1): 43–57.

Gard M. 2014. Modelling changes in salmon habitat associated with river channel restoration and flow-induced channel alterations. *River Research and Applications* 30: 40–44.

Hardy TB, Addley RC, Saraeva E. 2006a. Evaluation of instream flow needs in the lower Klamath River, Phase II, Final Report. Technical report, Institute for Natural Systems Engineering, Utah State University, Logan, UT.

Hardy TB, Shaw T, Addley RC, Smith GE, Rode M, Belchik M. 2006b. Validation of chinook fry behavior-based escape cover modeling in the lower Klamath River. *International Journal of River Basin Management* 4(2): 1–10.

Harvey BC, Railsback SF. 2009. Exploring the persistence of stream-dwelling trout populations under alternative real-world turbidity regimes with an individual-based model. *Transactions of the American Fisheries Society* 138: 348–360.

Hayes JW, Jowett IG. 1994. Microhabitat models of large drift-feeding brown trout in three New Zealand rivers. *North American Journal of Fisheries Management* 14(4): 710–725.

Heath MW, Wood SA, Brasell KA, Young RG, Ryan KG. 2015. Development of habitat suitability criteria and in-stream habitat assessment for the benthic cyanobacteria Phormidium. *River Research and Applications* 31: 98–105.

Holsinger L, Keane RE, Isaak DJ, Eby L, Young MK. 2014. Relative effects of climate change and wildfires on stream temperatures: a simulation modeling approach in a rocky mountain watershed. *Climate Change* 124: 191–206.

Hosking JRM. 1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)* 52(1): 105–124.

Isaak DJ, Peterson EE, Ver Hoef JM, Wenger SJ, Falke JA, Torgersen CE, Sowder C, Steel EA, Fortin MJ, Jordan CE, Ruesch AS, Som N, Monestiez P. 2014. Applications of spatial statistical network models to stream data. *Wiley Interdisciplinary Reviews: Water* **1**(3): 277–294.

Jowett IG. 2002. In-stream habitat suitability criteria for feeding inaga (galaxias maculatus). *New Zealand Journal of Marine and Freshwater Research* **36**(2): 399–407.

Lange K. 1999. *Numerical Analysis for Statisticians*. Statistics and Computing, Springer-Verlag: New York, NY.

Mukhopadhyay N. 2000. *Probability and Statistical Inference*. Marcel Dekker: New York, NY.

Newcomb TJ, Orth DJ, Stauffer DF. 2007. Habitat evaluation. In *Analysis and Interpretation of Freshwater Fisheries Data*, Brown ML, Guy CS (eds). American Fisheries Society: Bethesda, MD; 843–886.

Persinger JW, Orth DJ, Averett AW. 2011. Using habitat guilds to develop habitat suitability criteria for a warm water stream fish assemblage. *River Research and Applications* **27**(8): 956–966.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.

Royle JA, Chandler RB, Yackulic C, Nichols JD. 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution* **3**(3): 545–554.

Sandoval-Solis S, Teasley RL, McKinney DC, Thomas GA, Pati~no Gomez C. 2013. Collaborative modeling to evaluate water management scenarios in the Rio Grande basin. *Journal of the American Water Resources Association* **49**(3): 639–653.

Silverman BW. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

APPENDIX A

A.1 R code to evaluate fit of potential PDFs to depth variable

```
# Evaluate fit of potential PDFs to depth variable.
# ChinookFry is the name of the data object that contains the depth data.
# Define the data we'll be fitting a HSC curve to.
depth <- ChinookFry$Depth

# load the lmomco package
library(lmomco)

# Create a function that relies on the package "lmomco" to:
# 1: Estimate the parameters of a candidate PDF using observed data.
# 2: Calculate -log(likelihood) given the
#    parameter estimates and candidate distribution.
EstFitReport <- function(distn, dat){
X1 <- lmoms(dat)
parEst <- lmom2par(X1, type=distn)
CallLabel <- paste("pdf", distn, sep="")
nll <- (-sum(log(do.call(CallLabel, args=list(x=dat, para=parEst), quote=T))))
list(pars=parEst$para, nll=nll)
}

# Rayleigh distribution
EstFitReport("ray", depth) -LL=-591.5406
# Gamma distribution
EstFitReport("gam", depth) -LL=-636.5174
# Rice distribution
EstFitReport("rice", depth) -LL=-560.517
```

A.2 R code to scale PDF values between 0 and 1 and plot HSC curve.

```
# Scale PDF values so all HSC values lie between 0 and 1.
# Plot estimated curve to visually inspect fit to data.

# Create a function that extracts the parameter estimates a lmomco object
makePARA <- function(distn, ests){
para <- list()
para$type <- distn
```

```
para <- ests
para$source <- paste("par", distn, sep = "")
para
}


# Create a function that scales all HSC curve values to between 0 and 1,
# and plot the curve on a representation of the data, in this case a histogram.
plotScaledFit <- function(Range, distn, pars){
CallLabel <- paste("pdf", distn, sep = "")
newPARA <- makePARA(distn, pars)
vals <- do.call(CallLabel, args = list(x = Range, para = newPARA), quote = T)
max.val <- max(vals)
scaled.vals <- vals/max.val
list(values = scaled.vals, scaler = max.val)
}


# Create a histogram, but do not plot because we want the y-axis of the
# histogram to match the scale of our HSC values.
H <- hist(depth, breaks = "FD", plot = F)


# Adjust the histogram so the y-axis lies between 0 and 1.
H.nls <- H
H.nls$counts <- H$counts/max(H$counts)


# Create a dense set of numbers representing the range of depths
plotRange = seq(0, 1.5, 0.01)


# Plot histogram with axis labels
plot(H.nls, xlab = "Depth", main = "Chinook Fry", cex.lab = 1.5, cex.main = 2,
ylab = "Scaled Frequency")


# Add estimated HSC curve to histogram
lines(plotRange, plotScaledFit(plotRange, "gam",
    EstFitReport("gam", depth)$pars)$values, lwd = 3)
```

A.3 R code to create and add confidence intervals to HSC curve.

```
# Create a function that:
# Creates bootstrap confidence intervals of the PDF parameters,
# and the HSC curve values.
BootstrapCIs <- function(data, distn, nBoots, alpha, plotRange){
FitEst <- function(distn,data){
X1 <- lmoms(data)
parEst <- lmom2par(X1, type = distn)
parEst$para
}


makePARA <- function(distn, ests){
para <- list()
para$type <- distn
para$type <- ests
para $source <- paste("par", distn, sep = "")
para
}
```

```
plotScaledFitted <- function(Range, distn, pars){
CallLabel <- paste("pdf", distn, sep="")
para <- list()
para$type <- distn
para$para <- pars
para$source <- paste("par",distn,sep="")
newPARA <- para
vals <- do.call(CallLabel, args=list(x=Range, para=newPARA), quote=T)
max.val <- max(vals)
scaled.vals <- vals/max.val
scaled.vals
}

BootMat <- t(replicate(nBoots, FitEst(distn, sample(data,length(data),
        replace=T)),simplify="matrix"))

# Confidence Interval for parameter values
ParamCI <- apply(BootMat, 2, quantile, probs=c(alpha/2,(1-alpha/2)))
ParamSE <- apply(BootMat, 2, sd)

# Confidence Interval for curves
BootCurves <- apply(BootMat, 1, plotScaledFitted,Range=plotRange, distn=distn)
CurveCI <- apply(BootCurves, 1, quantile, probs=c(alpha/2,(1-alpha/2)))
Curve.CI.Lower = CurveCI[1,]
Curve.CI.Upper = CurveCI[2,]
list(Parameters.CI=ParamCI, Parameters.SE=ParamSE,
Curve.CI.Lower=Curve.CI.Lower,
Curve.CI.Upper=Curve.CI.Upper)
}

# Run function to obtain confidence intervals
# Note: may take several minutes to run
ConfInts <- BootstrapCIs(data=depth, distn="gam", nBoots=2000,
alpha=0.05, plotRange=seq(0,1.5,0.01))

# Plot histogram and fitted curve again
plot(H.nls, xlab="Depth", main="Chinook Fry", cex.lab=1.5, cex.main=2,
ylab="Scaled Frequency")
lines(plotRange, plotScaledFit(plotRange, "gam",
    EstFitReport("gam", depth) $pars)values, lwd=3)

# Add the upper and lower 95% confidence intervals
lines(ConfInts$Curve.CI.Lower~plotRange, lty=5, lwd=1)
lines(ConfInts$Curve.CI.Upper~plotRange, lty=5, lwd=1)
```
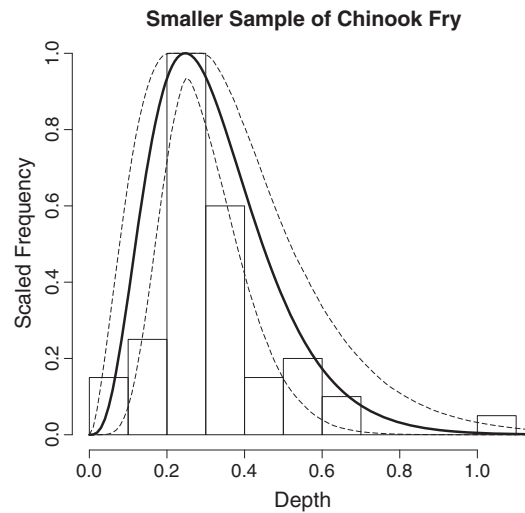
A. 4 Example of confidence intervals for smaller sample size

Figure A.1. Estimated habitat suitability criteria (HSC) curve (solid line) with 95% confidence intervals (dashed lines) for a randomly sampled reduced data set of 50 observations of the depth of habitats used by Chinook salmon fry in the Klamath River. The estimated HSC curve is generated by fitting the reduced data to a gamma distribution and then scaling the fitted PDF values by their maximum to restrict values between zero and one. The 95% confidence interval curves are generated by a non-parametric bootstrapping routine. This plot was generated to show how a smaller sample size will result in wider HSC curve confidence intervals with a less precise shape, as compared with the larger sample size revealed in the main text of the manuscript