Extrapolating Capacity Estimates to a Linear Stream Network

Kevin See

January 06, 2021

Contents

[1] Biomark, Inc.

Introduction

For several years, our development of a QRF capacity model has relied on extrapolation to master sample points to interpolate between CHaMP (Columbia Habitat Monitoring Program) sites and generate capacity estimates on larger spatial scales. Recently there has been interest in moving from a point-based prediction layer to a line-based prediction layer to aide with the interpretation and visualization of the capacity predictions. In this document, we present our method for doing so, and include some comparisons with the point-based estimates.

Methods

Master Sample Points

The master sample points were generated in the design phase of CHaMP. These 551,046 sites were selected from the NHD Plus 1:100,000 stream layer covering WA, OR and ID at an average density of one site per kilometer (Larsen et al. 2016). Each CHaMP site where direct QRF capacity estimates were made corresponds to one of these master sample points. CHaMP generated a number of of attributes for each master sample point, referred to here as globally available attributes (GAAs) because they are associated with every master sample point across all watersheds.

The original extrapolation model used the log of capacity estimates at each CHaMP site (fish / m) as the response, and various GAAs as covariates. The model was fit using the *svyglm* function in the *survey* (Lumley 2004) package with R software (R Core Team 2019), accounting for the various survey design weights within

Table 1: Attributes available at every master sample point, used as covariates in extrapolation model.

ShortName	Name	Description
TRange	Temeprature Range	Mean Temperature Range from PRISM data
Elev_M	Elevation	Elevation of site as extracted from the 10 m Digital Elevation Model
CHaMPsheds	CHaMP Watershed	CHaMP Watershed site falls in if appropriate
NatPrin1	Natural Class PCA 1	Natural Classification PCA 1 Score
DistPrin1	Disturbance Class PCA 1	Disturbance Classification PCA 1 Score
SrtCumDrn	Drainage Area (sqrt)	Square root of the cumulative drainage
StrmPwr	Stream Power	Stream Power
Slp_NHD_v1	Slope	Slope of Flowline (m/m) from the NHD Plus file
Channel_Type	Channel Type	Geomorphic Channel Type from Beechie Layer
WIDE_BF	Bankfull Width - modeled	Modeled bankfull width of stream, (m)
S2_02_11	Average August Temperature	NorWeST 10 year average August mean stream temperatures for 2002-2011

Table 2: Attributes available at every 200m reach, used as covariates in extrapolation model.

ShortName	Description
slope	Stream gradient
rel_slope	Relative slope. Reach slope minus upstream slope
Sinuosity	Reach sinuosity. 1=Straight, 1< sinuous
regime	Flow regime. 1= mixed, 2=snow dominated, 3=rain dominated.
alp_accum	Number of upstream cells in alpine terrain
fines_accu	Number of upstream cells in fine grain lithologies
flow_accum	Number of upstream DEM cells flowing into reach
grav_accum	Number of upstream cells in gravel producing lithologies
p_accum	Number of upstream cells weighted by average annual precipitation.
fp_cur	Current unmodified floodplain width
S2_02_11	NorWeST 10 year average August mean stream temperatures for 2002-2011
DistPrin1	Disturbance Classification PCA 1 Score
NatPrin1	Natural Classification PCA 1 Score
NatPrin2	Natural Classification PCA 2 Score

each CHaMP watershed. We then used that model to predict capacity at every master sample point that was not a CHaMP site. To roll up capacity estimates to larger spatial scales, the average predicted capacity of master sample points along a stream was multiplied by the length of that stream, and then combinations of streams could be added together to generate overall capacity estimates for a watershed.

Line Network

We adapted this method to using a stream layer created by Morgan Bond and Tyler Nodine at the Northwest Fisheries Science Center. This layer consisted of a line file divided into 200m reaches with various attributes attached to each reach. The line file is based on the National Hydrography Dataset High Resolution (NHDPlus HR) dataset, which has a higher resolution, 1:24,000, compared to the older layer that the master sample points were chosen from.

We determined which reach was closest to each CHaMP site, and used the predicted QRF capacity density of those CHaMP reaches as the response with the attributes attached to each 200m reach as covariates. We also took this opportunity to move to a random forest modeling framework. This accommodates possible non-linear

Table 3: Correlation coefficient between capacity estimates at the population scale using each method.

Species	Model	r
Chinook	CHaMP	0.934
Chinook	DASH	0.903
Chinook	Redds	0.908
Steelhead	CHaMP	0.866
Steelhead	DASH	0.990
Steelhead	Redds	0.986

or saturating effects of some of these covariates on capacity predictions, and prevents the extrapolation model from predicting capacity values well above or well below the range of predictions at CHaMP sites.

Range of Covariates

We examined the range of the covariates used in each method, and compared it to the range of values found at CHaMP sites or reaches. This exercise provides some context about how representative the suite of CHaMP sites are compared to the rest of the Columbia River Basin.

Capacity Comparisons

We computed the total capacity of each species in each population using both methods, for summer juveniles (using both CHaMP and DASH habitat metrics) and redds, and compared them. The correlations between the two estimates are shown in Table 3.

We plotted one estimate against the other in Figure 5, and showed the relative difference in Figure 6.

Maps

This shows the difference in how the results can be visualized.

Discussion

Extrapolations of QRF predictions are useful for higher-level spatial analyses or comparisons, such as at the watershed level. Examining predictions at individual master sample points or 200m reaches should be discouraged. For that scale, detailed habitat data should be collected, by using a protocol like DASH, and direct estimates of capacity can be made using a QRF model. On the other hand, extrapolation summaries of capacity at the watershed scale, for various species and life-stages, can be useful in broad prioritization discussions, to determine what life-stages and watersheds to target for rehabilitation.

For most of the GAAs, the range of values represented at CHaMP sites or reaches overlapped with the range of values in other places, with a few exceptions. The most notable is modeled precipitation (Precip) in the Clearwater basin (Figure 2. We did not use Precip as a covariate in the master sample points extrapolation model, but it does indicate that something about the conditions in the Clearwater may be different from other places with the interior Columbia River Basin, and therefore extrapolations to that area should be scrutinized carefully.

For both species, across all three QRF models, the two extrapolation models resulted in estimates of total capacity at the population scale that are very highly correlated. The linear network estimates were often greater than the master sample point estimates, to a greater or lesser degree, but not always.

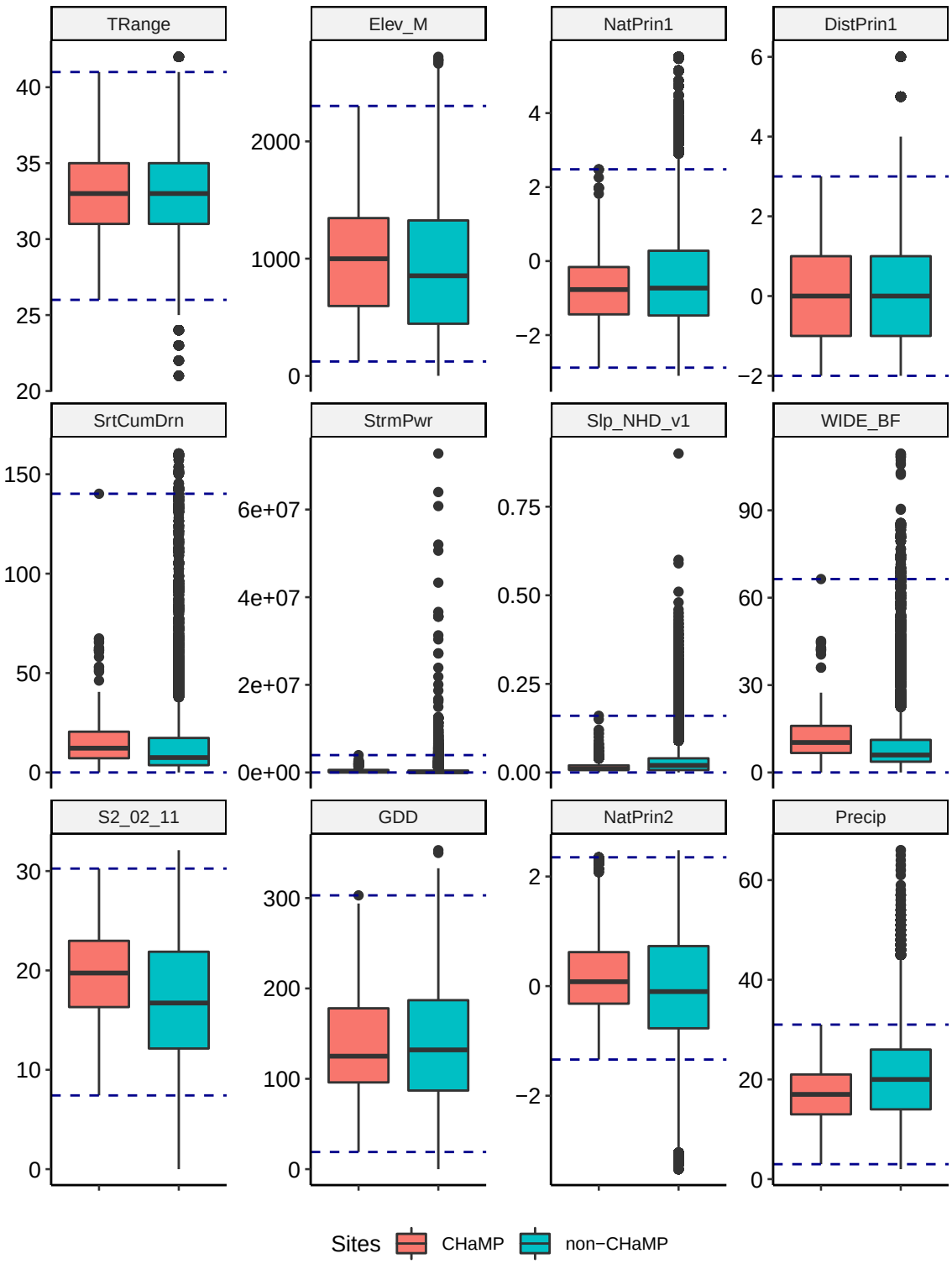

Figure 1: Boxplots of GAA values at CHaMP sites and non-CHaMP master sample points. Horizontal lines represent range of values at CHaMP sites.

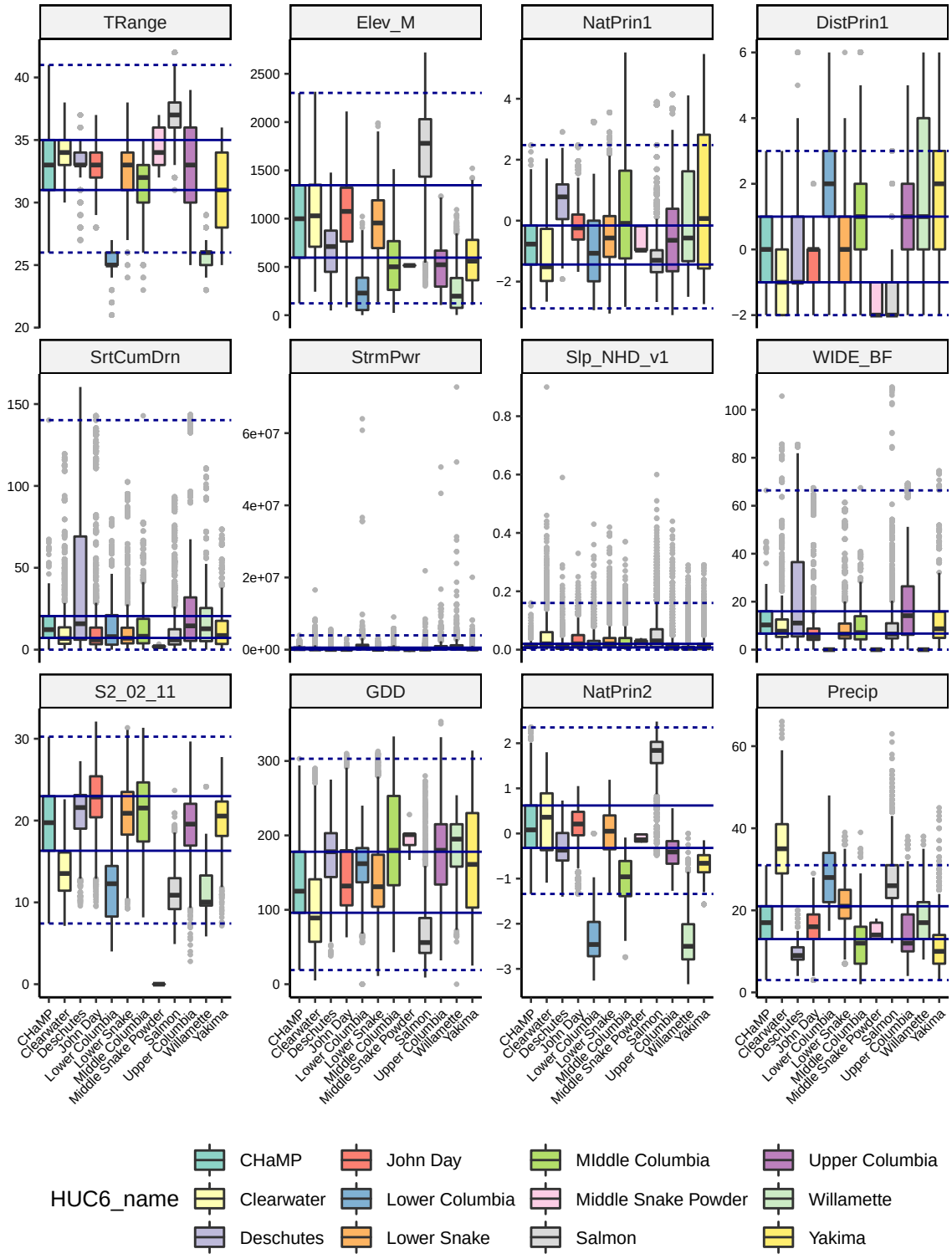

Figure 2: Boxplots of GAA values at CHaMP sites and non-CHaMP master sample points, colored by HUC6. Horizontal lines represent range of values at CHaMP sites (dashed) and the 25th and 75th quantiles of the CHamP sites (solid).

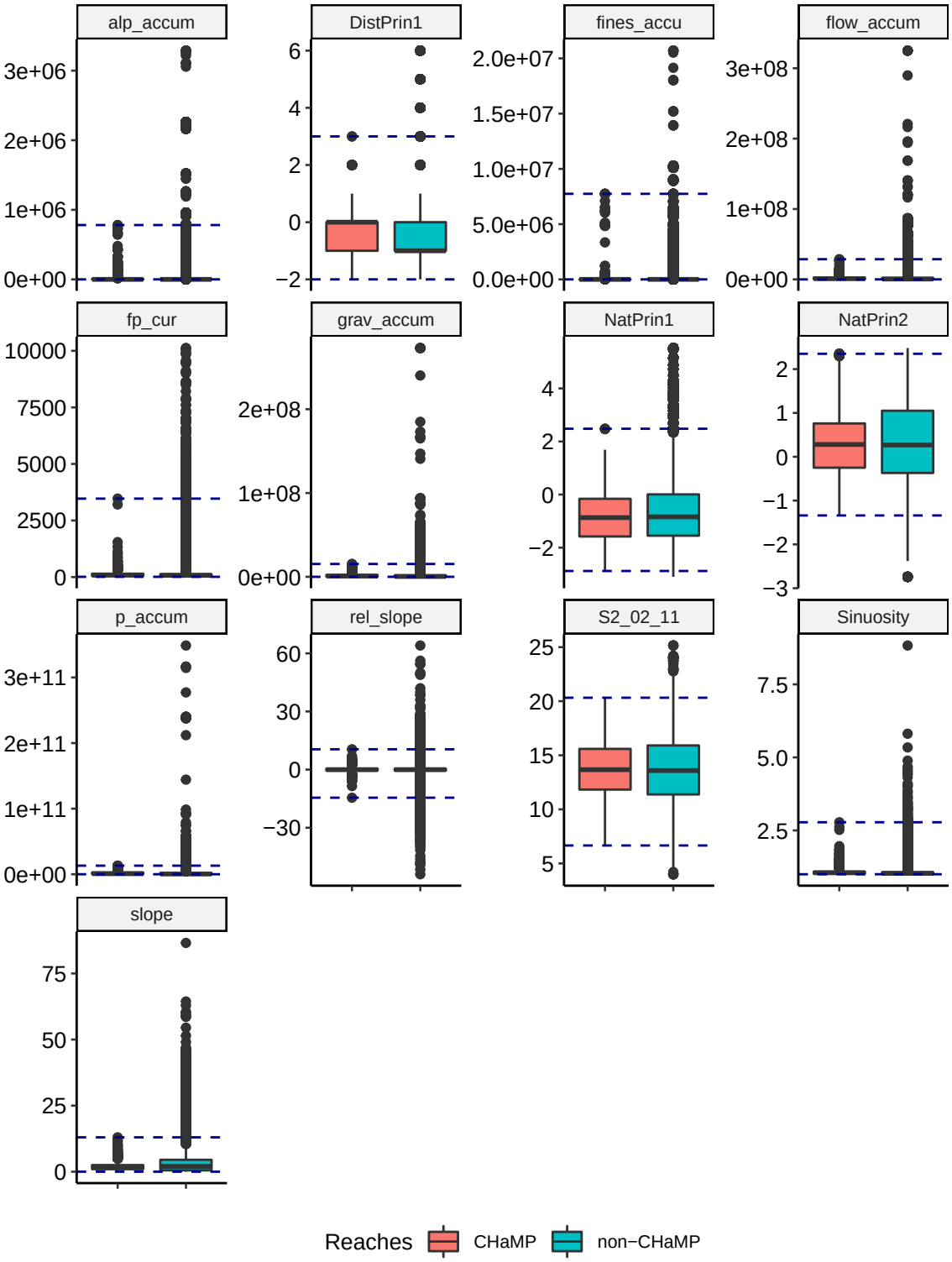

Figure 3: Boxplots of GAA values at CHaMP reaches and non-CHaMP 200 m reaches. Horizontal lines represent range of values at CHaMP sites.

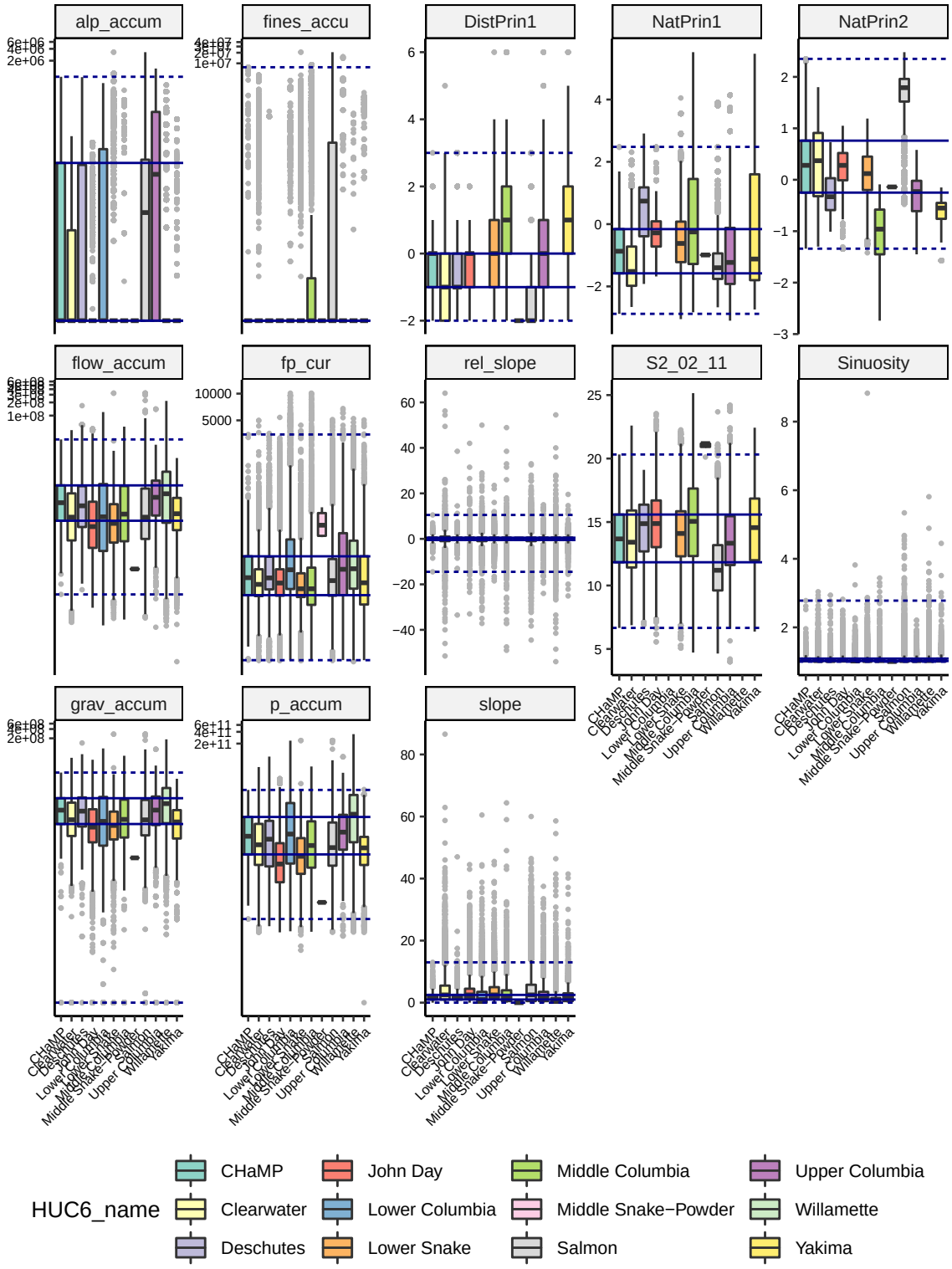

Figure 4: Boxplots of GAA values at CHaMP reaches and non-CHaMP 200 m reaches, colored by HUC6. Horizontal lines represent range of values at CHaMP sites (dashed) and the 25th and 75th quantiles of the CHamP sites (solid).

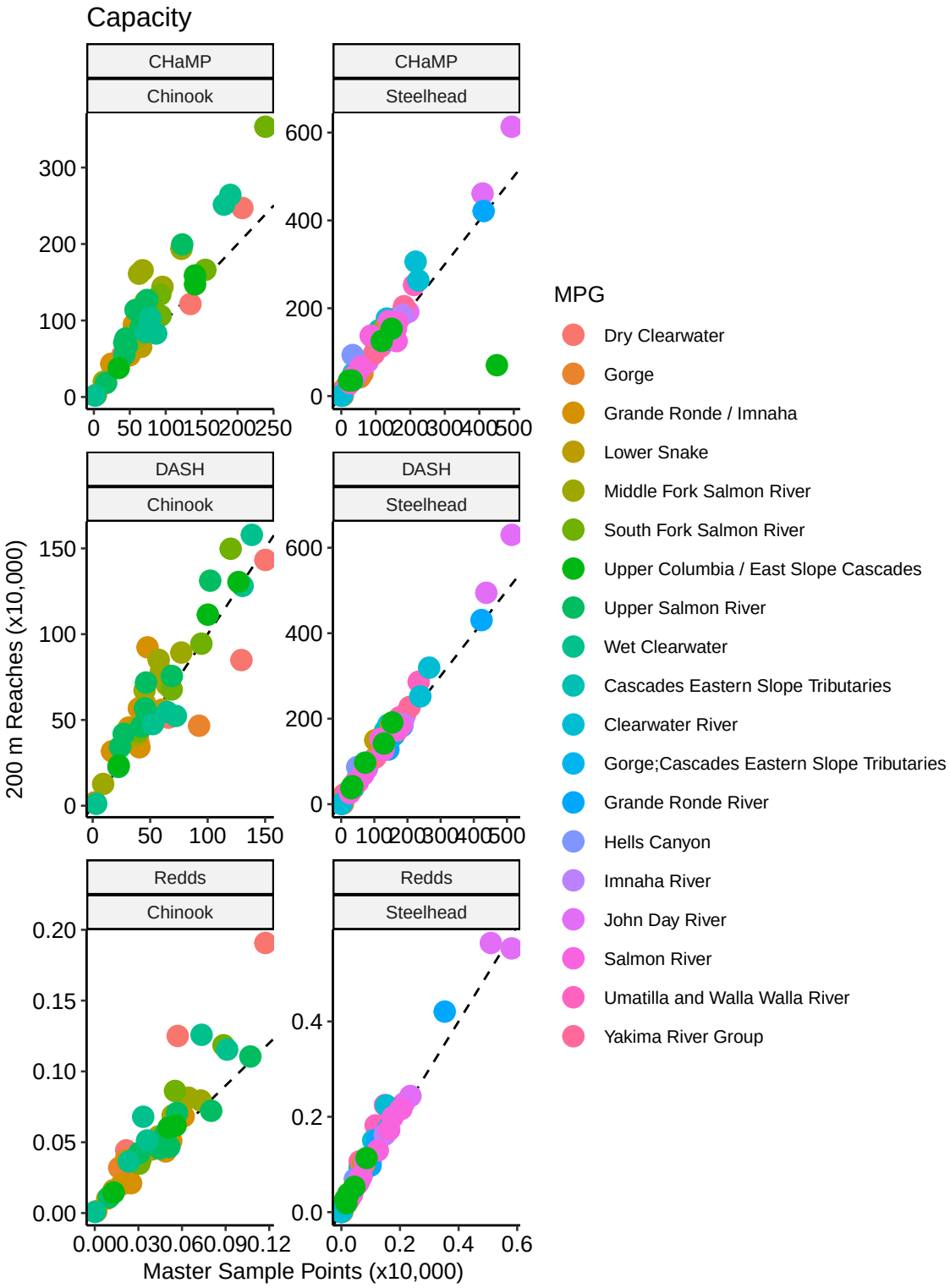

Figure 5: Capacity estimates for each population, calculated with the master sample points method on the x-axis and the line network on the y-axis.

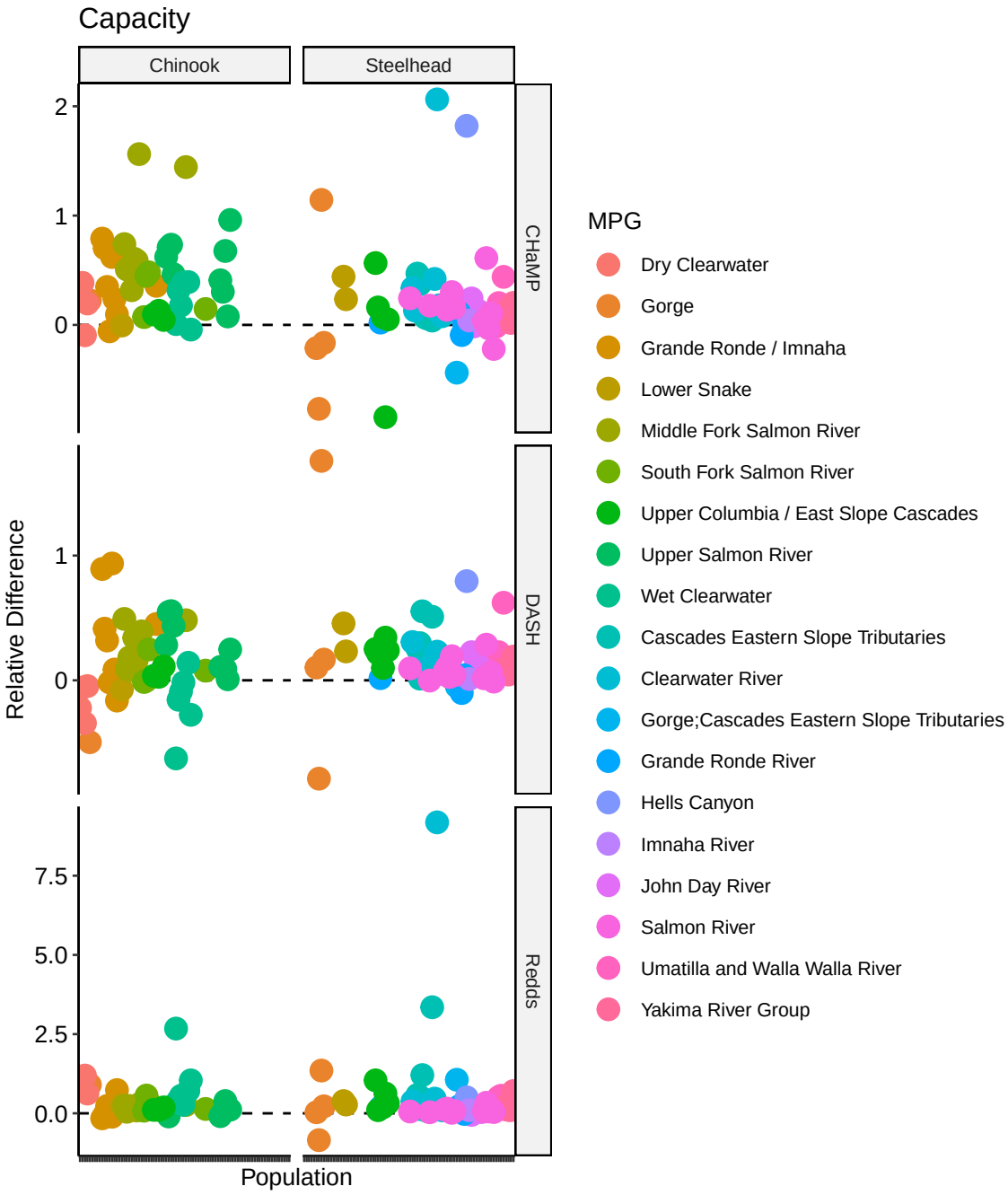

Figure 6: Relative difference between the capacity estimates for each population, using the master sample points method as the reference.

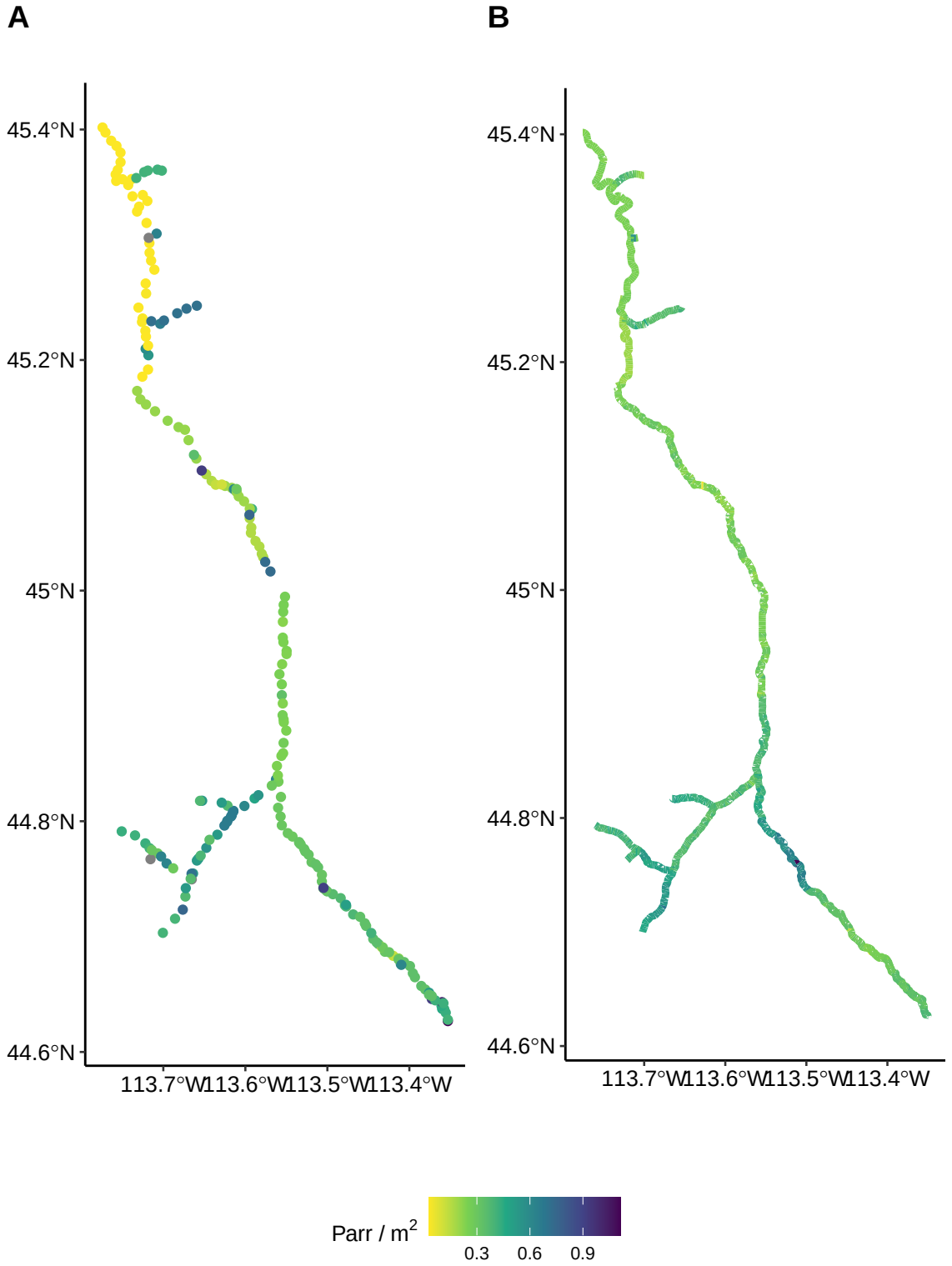

Figure 7: Plots of Chinook parr capacity in the Lemhi, using the master sample points method (A) and the 200 m reach method (B).

Changing the modeling framework from linear regression to a random forest has several benefits. Primarily, it provides a method to constrain extrapolation predictions naturally, even when the extrapolation covariates are beyond the range found at CHaMP sites. In addition, random forests accommodate potential non-linear associations between capacity predictions and GAAs while handling correlations among GAAs. The sample size of CHaMP sites with QRF predictions of capacity is sufficient to fit a random forest model, so we have no concerns about the "data-hungry" nature of this framework for this situation.

Although the master sample point method has been used for several years, there is no reason to believe estimates from that method are inherently superior to using a line network, so even in the cases when the two models result in different estimates of capacity, it is difficult to say which is "better". On the other hand, there are several reasons to support using the line network method, apart from the actual results, primarily based on the ease of interpretation. Extrapolation to a line network involves capacity predictions at actual 200 m reaches along a stream network, while the master sample point method provides estimates at instantaneous "points" on the landscape. The summation of capacity to larger spatial scales is more straightforward when using a line network, and the maps that can be created are easier to interpret (Figure 7).

Therefore, we conclude that the extrapolation to a linear network method presented here is superior to the master sample point method, and should be adopted moving forward for examining QRF outputs at large spatial scales.

References

Larsen, D. P., C. J. Volk, D. L. Stevens Jr, A. R. Olsen, and C. E. Jordan. 2016. An overview of the Columbia Habitat Monitoring Program's (CHaMP) spatial-temporal design framework. South Fork Research.

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

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