**Modeling CHaMP Metrics on Globally Available Attributes**

Matt Nahorniak, Carol Volk, Chris Jordan, Phil Larsen

December, 2015

**Introduction**

CHaMP metrics are measured at, typically, 45 points along a stream network within each CHaMP watershed, in a sampling design that includes both annual and 3 year rotating panel sites. The sampling design is optimized for estimation of the distribution (including mean and variance) of CHaMP metrics at the spatial scale of watershed, or at least over a large portion of a watershed. However, importance is also being placed on making spatially continuous estimates of CHaMP metrics at all points along stream networks of interest, throughout the Interior Columbia River Basin (ICRB), rather than simply at CHaMP watershed spatial levels. Nodeled estimates enable generation of maps showing estimated CHaMP distributions in a spatially continuous manner, providing an additional tool for habitat management. In addition, modeled estimates can be used to augment design based estimates, where limited or no sampling data exists, in order to provide inputs to habitat and salmonid life cycle models.

Fitting models of CHaMP metrics on globally available attributes (GAA) is being done to estimate CHaMP metrics continuously at every point along stream networks of interest. Model based estimates for selected CHaMP metrics are being made both within current CHaMP watersheds, as well to several watersheds where CHaMP data is not being taken (including the South Fork Clearwater, Lower Clearwater, Lolo, Lochsa, and Upper Salmon River Tributaries Above Redfish Lake)

For successful, informative models capable of both predicting CHaMP metrics within CHaMP watersheds and extrapolating CHaMP metrics into non-CHaMP watersheds, at least two criteria must be met: 1) For the GAAs that are used as covariates in the models, the range of GAAs observed at measured CHaMP sites must cover the range of GAAs at all sites at which we intend to make predictions from the model. We refer to this as “GAA Coverage”, and it includes both numeric and categorical GAA variables. 2) A subset of all GAAs available must be statistically informative to the CHaMP metric being validated. In other words, there must be observable relationships between our GAAs and CHaMP metrics. This is examined through cross validation at various spatial levels.

CHaMP metrics for which models have been attempted to date include Percent Substrate < 2mm, Percent Substrate < 6mm, Sinuosity, SubD50, PoolResidDpth, DpthThlwg\_UF\_CV, and SLowerWater\_Pct; as well as the higher level CHaMP products of habitat suitability index (HSI) for juvenile and spawning steelhead and chinook, and net return on energy investment based capacity (NREI).

**Methods**

Simple multiple regression models have been constructed relating our CHaMP metrics of interest to available globally available attributes (GAA). Because CHaMP data are from a non-uniform probability GRTS [1] sampling design, it is necessary to take into account sample inclusion probabilities in the fitting of regression models [2]. To do this, we use model assisted regression [3], utilizing the R function *svyglm* from the package *survey* [4].

GAAs used to model each CHaMP metric (Table 1) were selected using forward stepwise regression with the function *add1*. Minimization of AIC was the criterion used as the criterion for inclusion, and covariates are include in models if they reduce total model AIC by value greater than about 3-5. Note that because GAAs selected may be correlated with one another, and because we are dealing with observational data, extreme caution should be made in interpreting model coefficients. Our intent is not to infer meaning from model coefficients, but rather to develop predictive models only.

Plots of measured vs predicted values for each CHaMP model are produced to visually assess model goodness of fit (Figure 1 and 2). Inspection of model residuals is performed to determine the need for a transformation of the metric being modeled, or other potential problems with the model fit (heteroscedasticity, non-linearity, etc.). If transformations are deemed necessary, then the metric being modeled is transformed and the entire process of variable selection and model fitting is repeated.

Table 1. Globally available attributes (GAA) used as covariates for extrapolation models



Figure 2. Measured vs predicted SubD50



Figure 1. Measured vs predicted juvenile steelhead wetted usable area (WUA) per meter of stream length

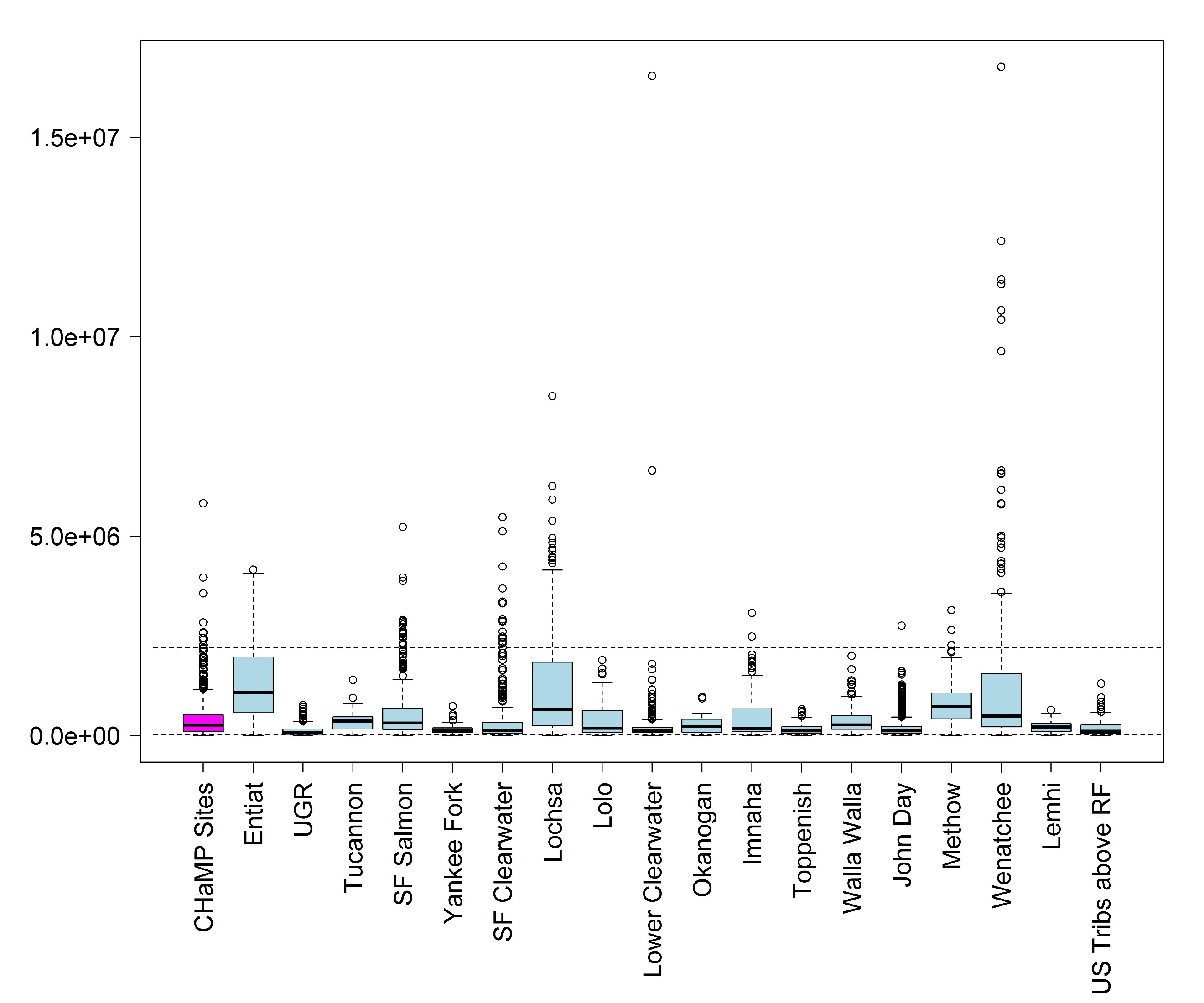
**Coverage**

GAAs selected for model inclusion include both continuous and categorical variables. In both cases, it is desirable that the distribution of GAA values covered across sites also included in the actual CHaMP sample be reasonably representative of the distribution of GAAs values included in the spatial region to which we want to apply these models. Because CHaMP sampling uses a robust survey design [ref TBD], coverage is generally excellent within the CHaMP spatial domain. However, coverage may be less representative where modeling is applied outside the CHaMP sampling domain. For categorical variables, meeting this requirement is often more difficult as 100% of the categorical levels included in the spatial domain being modeled must also be represented in the subset of sites actually sampled by CHaMP (and preferably a healthy sample size examples of each categorical levels are included in the CHaMP sample). This generally excludes, for the possibility of model inclusion, categorical variables for which more than a few levels exist. For example, many unique River Styles are present in the Interior Columbia spatial domains into which we want to extrapolate, and many of those unique River Styles do not exist at measured CHaMP sites. So River Style is not useful for extrapolation modeling. However, many of the continuous response parameters (sinuosity, valley confinement) or lower level categorical variables (valley class) that may feed into River Styles are, potentially, useful GAAs.

Histograms comparing the distribution of GAAs in sampled CHaMP sites to all sites within the ICRB where we intend to apply our models (Figure 3) are examined to visually assess the range of GAA variability covered by the CHaMP sample. Box and whisker plots (Figure 4) further break down the examination of coverage so coverage by individual watershed can be examined.

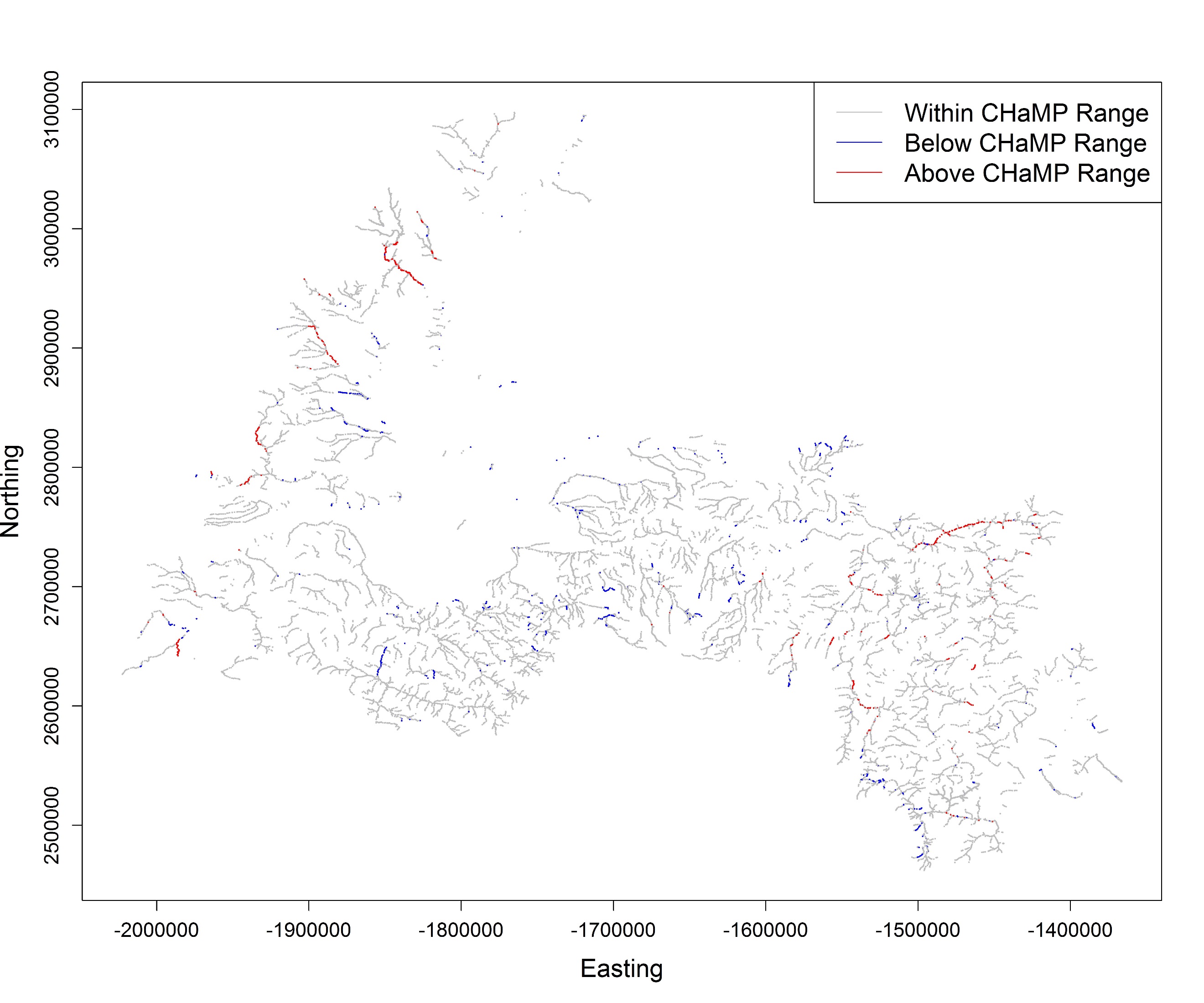
Figure 3. Distribution of GAA Stream Power in sampled CHaMP Sites (cyan) compared to the distribution of GAA BFW\_M in all ICRB watersheds of interest (red)



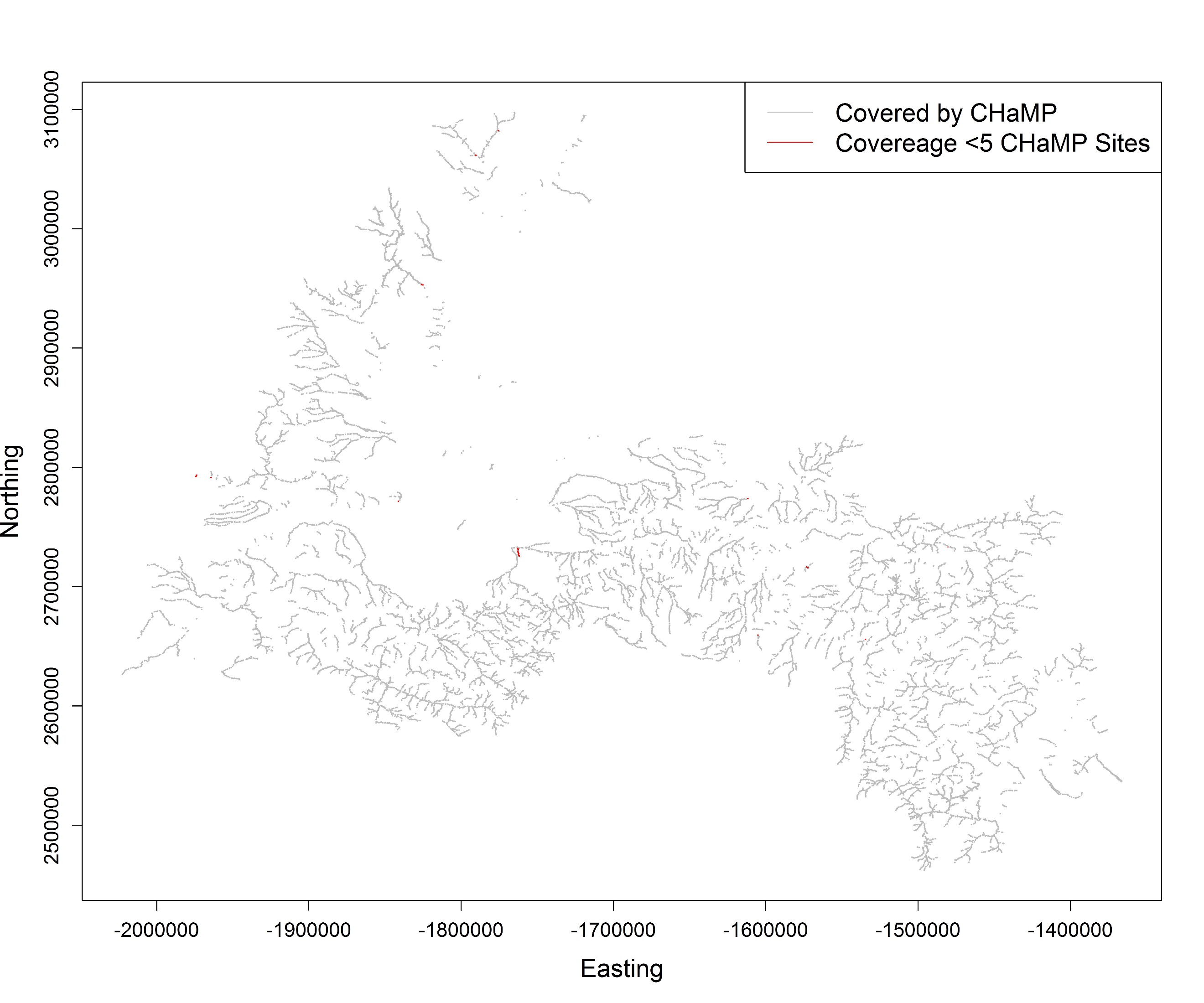
Figure 4. Coverage for GAA Stream Power in the ICRB, by watershed.

Typically, a few sites on the extreme tail of a distribution fall outside of the range of the CHaMP sample. Stream Power in the Wenatchee, for example, has a small number of sites with stream power above the maximum stream power observed in CHaMP sites

We also generate maps showing locations of interest within the ICRB where adequate coverage exists (Figure 5). We defined adequate coverage as all values of a GAA that fall within the 95% inter-quartile range of the values as measured in CHaMP sites. For stream power, we typically lack adequate coverage only in the very upper reaches of a watershed where stream power falls below the measured IQ range of measured CHaMP sites, or lower down in some main stem reaches where stream power is greater than the IQ range for measured CHaMP reaches.

Figure 5. Coverage for GAA Stream Power in the ICRB. Red and blue indicate BFW\_M values that are above and below the 97.5% and 2.5% quantiles, respectively, of Stream Power in measured CHaMP sites.

For categorical variables, we consider coverage to be adequate at a given ICRB location if, within the CHaMP sample, we have at least five sampled CHaMP sites at the categorical level of that location. For example, Channel Type (with categorical levels of: braided, cascade, confined, island braided, meandering, plane-bed, pool-riffle, step-pool, and straight) is used as a covariate in several models, and we find excellent coverage throughout the region of interest (Figure 6).

Figure 6. Coverage for categorical GAA Channel Type in the ICRB. Red indicates sites for which we have fewer than five CHaMP sites at the given level of Channel Type

**Model Validation**

Selecting a subset of model covariates from a large population of potential candidates invariably leads to an over-estimation of model precision if the modeler relies on r-squared or other residual based goodness of fit metrics [ref TBD]. Therefore, additional model validation is necessary to accurately assess goodness of fit and estimate the precision of predicted values.

For this assessment we attempt to validate models using three methods of differing spatial scales. First, we perform leave one out cross validation where one site at a time is removed during the model fitting process, and the model fit from the rest of the data is used to predict the remaining point. This is repeated for all points, such that an error (predicted value minus measured value) is obtained for each point. We then compare measured to predicted values at all points (Figures 7 and 8) and observe the distribution of error. The distribution of error is then used to calculate prediction uncertainties and r-squared values, reflective of the likely accuracy of using the models to predict points within currently sampled CHaMP watersheds.

Figure 8. Cross validation results, measured vs predicted D50 (R-squared = .53)



Figure 7. Cross validation results, measured vs predicted juvenile Steelhead wetted usable area per meter of stream length (R-squared = .62)



Because we also want to use CHaMP to predict outside of CHaMP watersheds, we need an estimates of error levels likely when doing this. The first method for this is to again use a leave-one-out cross validation method, except for this we leave data from an entire watershed out during each iteration of the model fitting stage. The fit model is then used to predict results at all measured points within the left out watershed (Figures 9 and 10). Because an entire watershed is left out of the process, we might expect the performance of the model in cross validation to be somewhat less than during cross validation within measured watersheds. This is the level of model performance we expect in non-CHaMP watersheds if we can assume that CHaMP watersheds are representative of non-CHaMP watersheds into which we predict points. Note that CHaMP watersheds were not selected at random from the broader population of IC watersheds, so this may not be a valid assumption.

Figure 10. Watershed level cross validation results for D50 (R-squared = .46)



Figure 9. Watershed Level Cross Validation Results for HSI Juvenile Steelhead capacity per meter (R-squared = .39)



Finally, and also to estimate error rates when extrapolating our model in watersheds without CHaMP data, we use external, non-CHaMP data. PIBO data [ref] is an external data source from a monitoring program which has some metrics in common with CHaMP metrics. We use our models to predict CHaMP metrics in non-CHaMP watersheds, at locations where PIBO has taken measurements (Figure 11). Since PIBO and CHaMP have a subset of metrics in common, we can use this to assess extrapolation performance. However, we can only do this for metrics that CHaMP and PIBO (or other external data sources) have in common. Higher level CHaMP metric (HSI and NREI metrics) cannot be cross-validated in because of this. Furthermore, we must recognize that CHaMP protocols do not always match, exactly, PIBO (or other data source) metrics, because of differences in training, calibration, protocols, etc.; so there may be a limit to how well these metrics cross validate even in the ideal case of an error free extrapolation model). Actual error rates are a combination of method-method (CHaMP-PIBO) noise and extrapolation model error.

Figure 11. Cross validation w/ PIBO data for D50 (R-squared = .25) 

Performance of extrapolation models varies widely by CHaMP metric (Table 2). HSI and NREI based capacity metrics generally perform well, as these metrics appear to be highly influenced by stream size (width, flow rate, etc.) which are highly informed by available GAAs. SubD50 (median particle size) and DpthThlwg\_UF\_CV showed fair performance in both within watershed and cross-watershed validation, and are likely to yield informative predictions external to CHaMP watersheds where CHaMP watersheds are representative. Models for PoolResidDth, LWFreq\_BF, and SlowWater\_PCT have some ability to predict within CHaMP watersheds, but should not be expected to perform well outside of CHaMP watersheds. Models for SIN, SubLT2, and SubLT6 did not perform well, even within CHaMP watersheds.

Table 2. Cross validation results for CHaMP metrics modeled, at three levels of cross-validation: individual site cross validation within modeled watersheds, leaving multiple watersheds out of model fitting process, and using external data for cross-validation (where available).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CHaMP Metric** | **Leave One Out Cross Validation R-squared estimate** | **Leave Multiple Watersheds Out Cross Validation** | **PIBO Comparison in Clearwater & Upper Salmon** | **Spatial Scale used for extrapolation model** |
| HSI\_CH\_Juv | 0.76 | 0.57 | N/A | All CHaMP Watersheds |
| HSI\_CH\_Spawn | 0.54 | 0.33 | N/A | All CHaMP Watersheds |
| HSI\_ST\_Juv | 0.62 | 0.39 | N/A | All CHaMP Watersheds |
| HSI\_ST\_Spawn | 0.61 | 0.37 | N/A | All CHaMP Watersheds |
| NREI | 0.63 | 0.63 | N/A | All CHaMP Watersheds |
| SubD50 | 0.53 | 0.43 | 0.25 | All CHaMP Watersheds |
| DpthThlwg\_UF\_CV | 0.52 | 0.33 | N/A | All CHaMP Watersheds |
| PoolResidDpth | 0.30 | 0.23 | 0.00 | All CHaMP Watersheds |
| LWFreq\_BF | 0.27 | 0.00 | 0.00 | All CHaMP Watersheds |
| SLowerWater\_Pct | 0.19 | 0.02 | 0.00 | All CHaMP Watersheds |
| SIN | 0.07 | 0.00 |  | YF-Tuc-UGR-Entiat |
| SubLT2 | 0.05 | 0.00 |  | All CHaMP Watersheds |
| SubLT6 | 0.16 | 0.00 |  | All CHaMP Watersheds |

As additional GAAs are generated or derived from River Styles, we will continue to refine and improve our models.

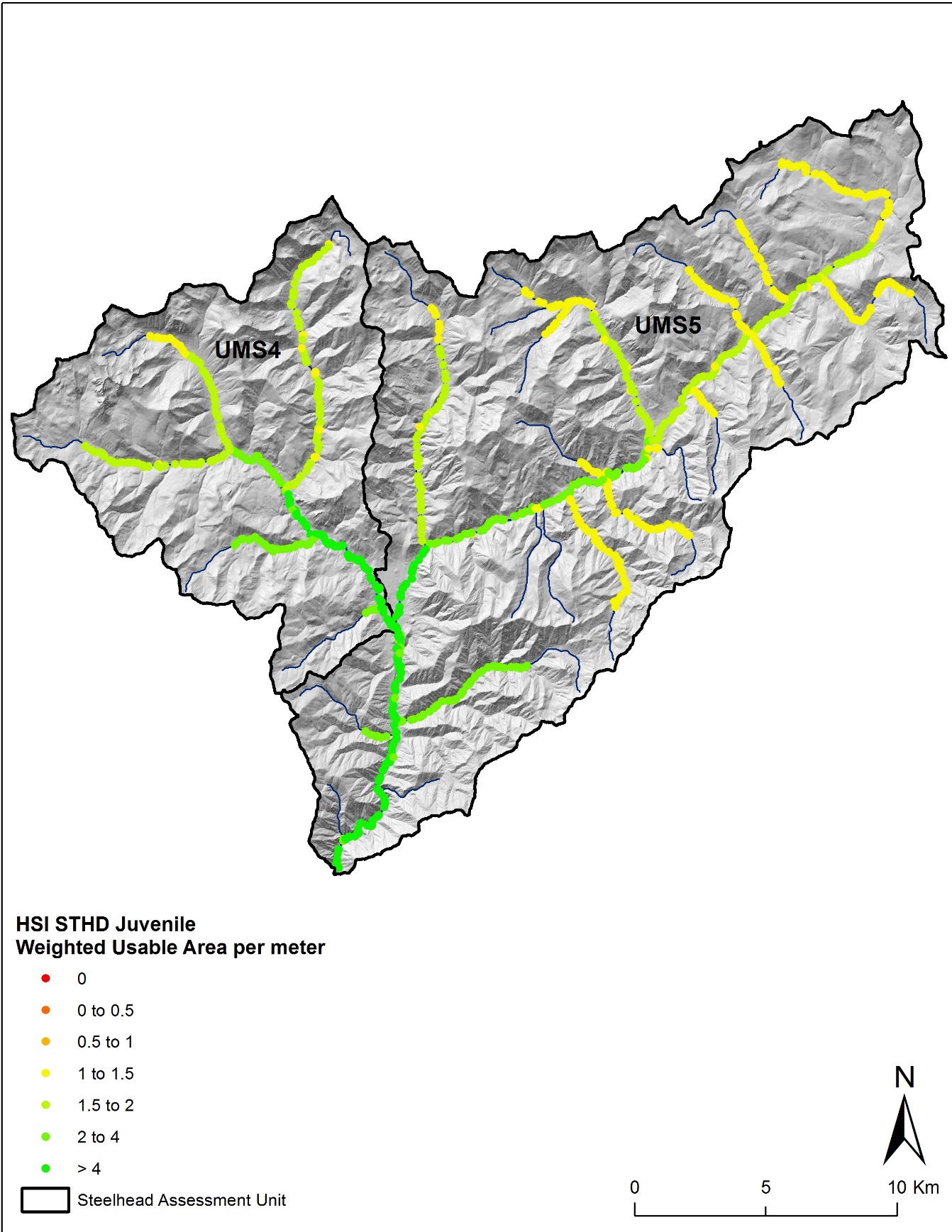
**Spatial Autocorrelation**

Note that in our model based estimation process, we do not incorporate spatial autocorrelation into our models. CHaMP sampling is performed using a GRTS [1] design, which generates spatially balanced samples. While these spatially balanced samples generally provide for greater efficiency in the estimation population distribution parameters than conventional random sampling, the samples generally do a poor job of estimating spatial autocorrelation, because sampled sites are rarely close together in space. Thus our samples generally do not have the information required to take advantage of spatial autocorrelation when predicting CHaMP metrics at sites near measured CHaMP sites.

**Map Generation**

One of the primary uses model based estimation of CHaMP metrics is the generation of continuous maps (Figure 12). GAA values are available at points forming a quasi-continuous continuum across the network (where continuous, in this case, mean available at approximately 1 km increments across the stream network), and for each point where GAAs are available, we can make a prediction for the level each CHaMP metric modeled. Care must be taken when interpreting such maps, as the maps themselves fail to display the level of uncertainty in the predictions; in addition, one should not assume that errors are spatially random. Maps must be viewed in conjunction with some assessment of the underlying model performance. In addition, extra caution must be made if predictions are made in areas for which the CHaMP sampling is non-representative.

Figure 12. Spatially continuous modeled estimates of juvenile steelhead wetted usable area per meter of stream length for the Yankee Fork watershed.



**Discussion**

For some CHaMP metrics, particularly the higher level capacity metrics from habitat suitability index (HSI) and net return on energy investment (NREI) models, we have models that we expect to be effective when used to predict values at sites not sampled by CHaMP within the ICRB region of interest. Other models (D50, DpthThlwg\_UF\_CV) show promise but could benefit from further refinement using more informative GAAs. For some models, however, we have little or no ability to predict outside of measured CHaMP sites and require significantly more informative GAAs to yield valid predictions.

**References**

1. Stevens Jr DL, and Olsen AR (2004). Spatially balanced sampling of natural resources. J Am Stat Assoc, 99:465, 262-278, DOI: 0.1198/016214504000000250
2. Nahorniak M, Larsen DP, Volk C, Jordan CE (2015) Using Inverse Probability Bootstrap Sampling to Eliminate Sample Induced Bias in Model Based Analysis of Unequal Probability Samples. PLoS ONE 10(6): e0131765. doi:10.1371/journal.pone.0131765
3. Lumley T (2004). Analysis of complex survey samples. J Stat Softw 9(1): 1-19.
4. Lumley T (2012). Survey: analysis of complex survey samples. R package version 3.28-2.