Revised QRF Habitat Capacity Model and RF Extrapolation Model Brief

Bryce N. Oldemeyer1,✉, and Mark Roes2,✉

February 14, 2024

1 Mount Hood Environmental, 803 Monroe St, Suite 106, Salmon, Idaho, 83467, USA  
2 Mount Hood Environmental, 39085 Pioneer Boulevard #100 Mezzanine, Sandy, Oregon, 97055, USA

✉ Correspondence: [Bryce N. Oldemeyer <[bryce.oldemeyer@mounthoodenvironmental.com](mailto:bryce.oldemeyer@mounthoodenvironmental.com)>](mailto:bryce.oldemeyer@mounthoodenvironmental.com), [Mark Roes <[mark.roes@mthoodenvironmental.com](mailto:mark.roes@mthoodenvironmental.com)>](mailto:mark.roes@mthoodenvironmental.com)

# Background

Quantile random forest (QRF) models have become popular for quantifying freshwater habitat carrying capacity due to their flexible framework that avoids common pitfalls associated with noisy data, correlated variables, and non-linear relationships. Recently, three QRF models were fit with fish-habitat data from fish observation studies and the Columbia Habitat Monitoring Program (CHaMP) and used to estimate habitat carrying capacity for ESA-listed populations of Chinook salmon and steelhead during three critical life-stages (juvenile summer parr, juvenile winter presmolt, and adult redds) for wadable streams within the Columbia River Basin. Model covariates were selected from >100 habitat metrics and chosen for their high predictive power (Appendix B of Idaho OSC Team, 2019; See et al. 2021). Since then, additional emphasis has been placed on the utility of the QRF capacity models to inform restoration project design and monitoring and increase the spatial extent of fish-habitat data using streamlined protocols (DASH - Carmichael et al. 2019). Therefore, we conducted a revised covariate selection process for the QRF models that prioritized 1) predictive power, 2) compatability with future DASH data collection, 3) informing restoration project development and monitoring, 4) minimal imputation for missing CHaMP data, and 5) low covariate correlation. Additionally, we evaluated the assumption made during initial QRF model development that a single model was appropriate for both Chinook salmon and steelhead during each of the three life stages.

Similarly, a random forest (RF) extrapolation model was used to predict habitat capacity across larger spatial scales where CHaMP and/or DASH data weren’t available (Appendix B of Idaho OSC Team). We revisited the globally available attributes (GAAs) included in the original RF extrapolation model and made minor modifications to the model that maintained covariates with high predictive power and included metrics that better aligned with the revised QRF model. To evaluate the differences between the original and revised QRF/RF models, we compared watershed carrying capacity estimates produced by the both sets of models for eight watersheds located within the Upper Salmon River basin.

This process resulted in revised QRF and RF extrapolation models that were more informative for restoration design and monitoring, included covariates that could be calculated using newly developed stream habitat protocols, and maintained a similar level of predictive power as the original models.

# Revised QRF Habitat Capacity Model

## Covariate selection process

Habitat covariates for the QRF habitat capacity models were generated from the CHaMP dataset or obtained from other publicly available sources (e.g. NorWest stream temperature data). In total, 129 habitat metrics were examined in the selection process. Covariates were aggregated into eleven metric categories and 1-4 covariates were chosen from each category based on following criteria:

1. What was the strength between the covariate and the response variable based on the maximal information coefficient (MIC) value?
2. Could the covariate be calculated using DASH data?
3. Was the covariate informative for restoration efforts?
4. How much data were missing and/or the amount of “0”s for the covariate in the fish-habitat dataset?
5. How correlated was the covariate with other covariates within the same metric category, particularly covariates with higher MIC scores?

Below is a simplified, theoretical example of how a covariate might be selected for a model.

–

*In the original QRF model, discharge was included as a covariate because it had a high MIC score and it made biological sense (i.e. discharge is a significant factor impacting fish habitat use and, presumably, habitat carrying capacity). Unfortunately, discharge isn’t that informative for restoration efforts because most restoration actions can’t create water. Discharge, like many habitat metrics, is highly correlated with other potential covariates which may have been left out of the original QRF model for any number of reasons (highly correlated with other model covariates, excluded to avoid overfitting, etc.). Using the revised model selection criteria, we observed that average thalweg depth has a MIC score nearly as high as discharge, is informative for restoration efforts, can be calculated from DASH, and is highly highly correlated with discharge. Based on all the information above, mean thalweg depth would be substituted for discharge in the model.*

–

The covariate selection process was conducted independently for both species for all three life stages to test the assumption made during the original QRF model development that it was appropriate to apply the same life stage models to both species.

## Covariate selection results

There were 12-14 covariates selected for each of the six QRF habitat capacity models. While the relative importance of the final covariates in the three life stage models differed between species, the final covariates themselves were nearly identical. (Figure 1 , Figure 2, and Figure 3 ). This confirmed that one model for both species per life stage was appropriate. Therefore, we consolidated the species-specific models into a single winter juvenile, summer juvenile, and redd models. (Table 1). Examination of covariate partial dependence plots from the revised QRF habitat capacity models indicated effects that were generally biologically intuitive and can be found in Section ??

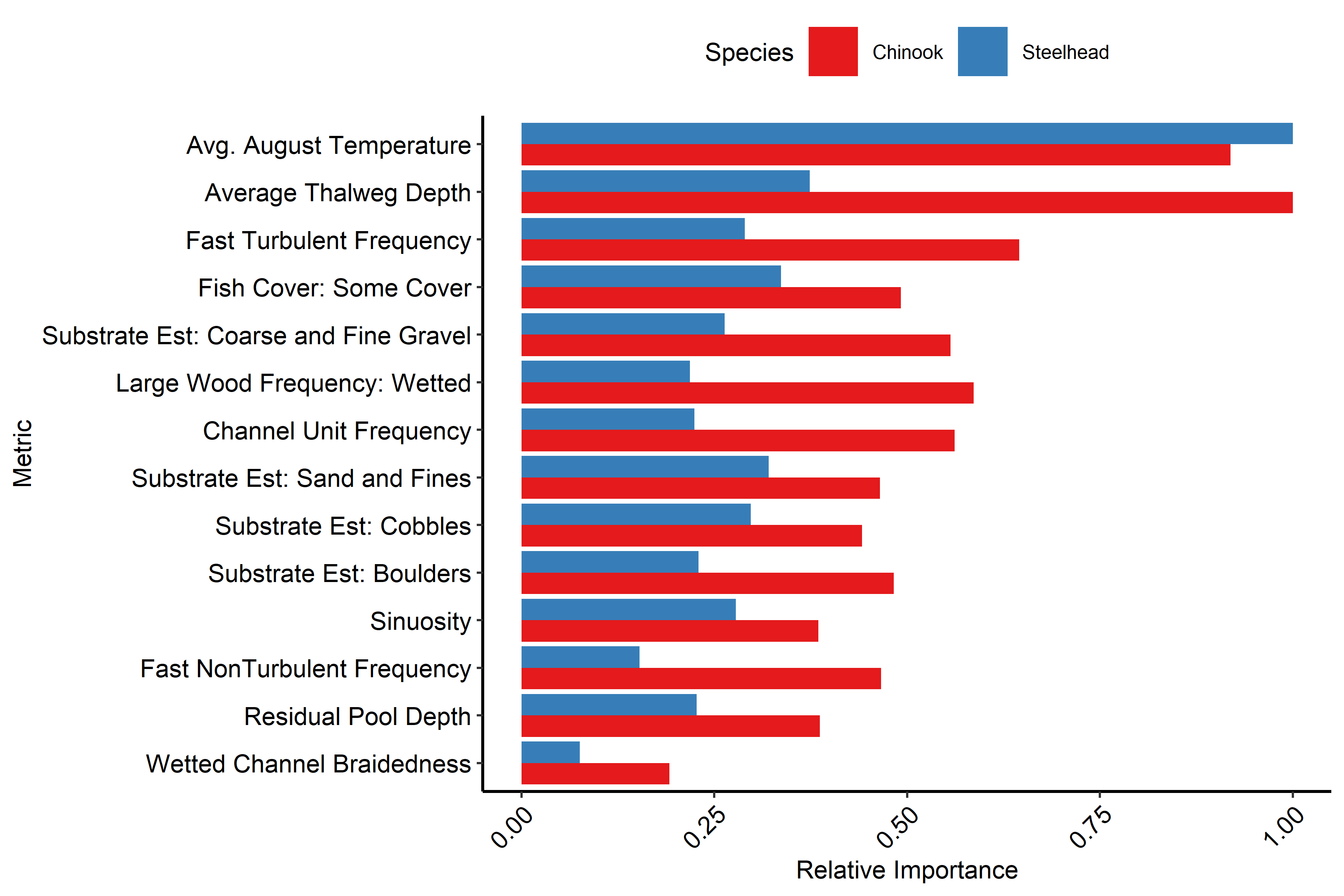


Figure 1: Relative importance plots for covariates included in the revised juvenile summer QRF models

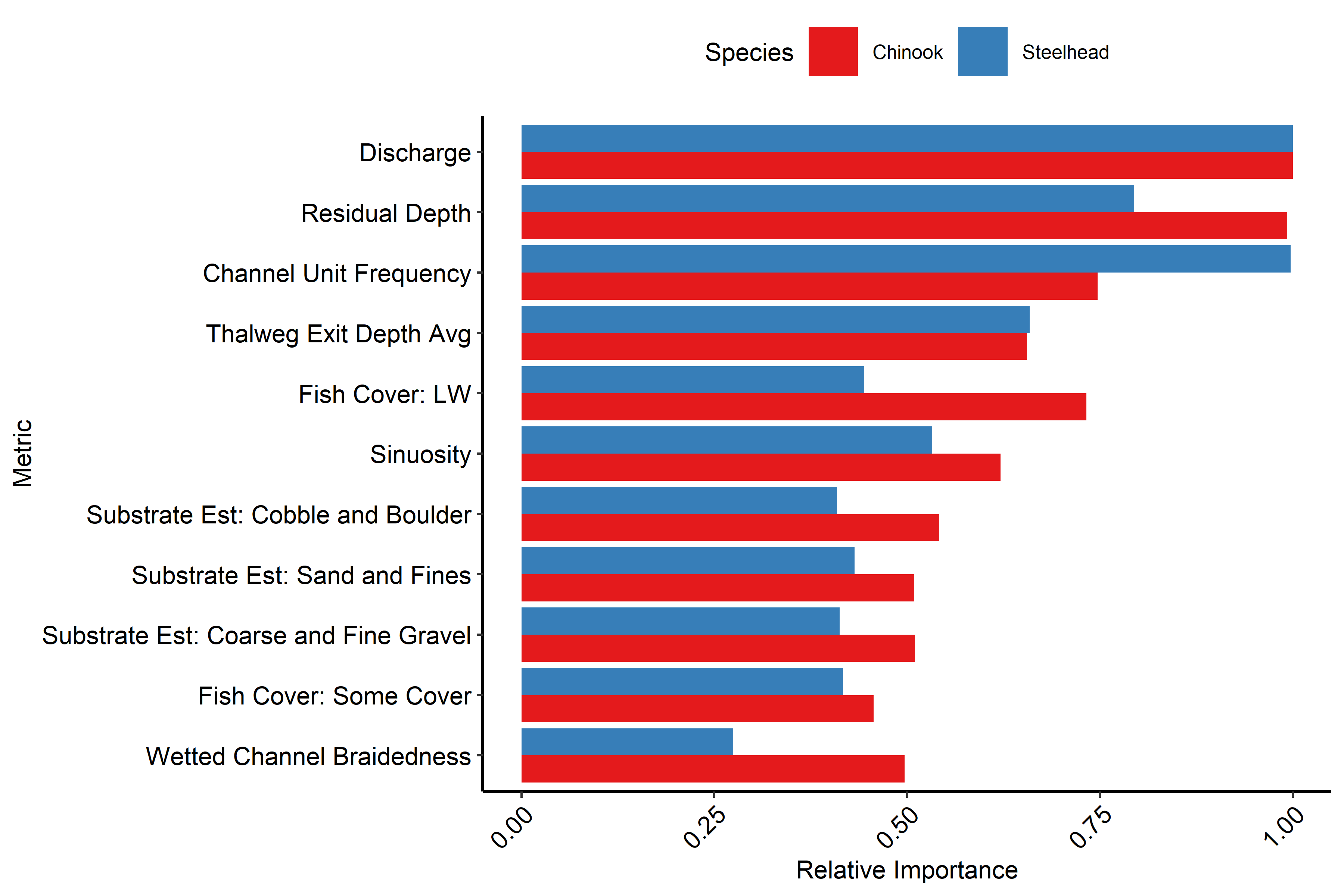


Figure 2: Relative importance plots for covariates included in the revised juvenile winter QRF models

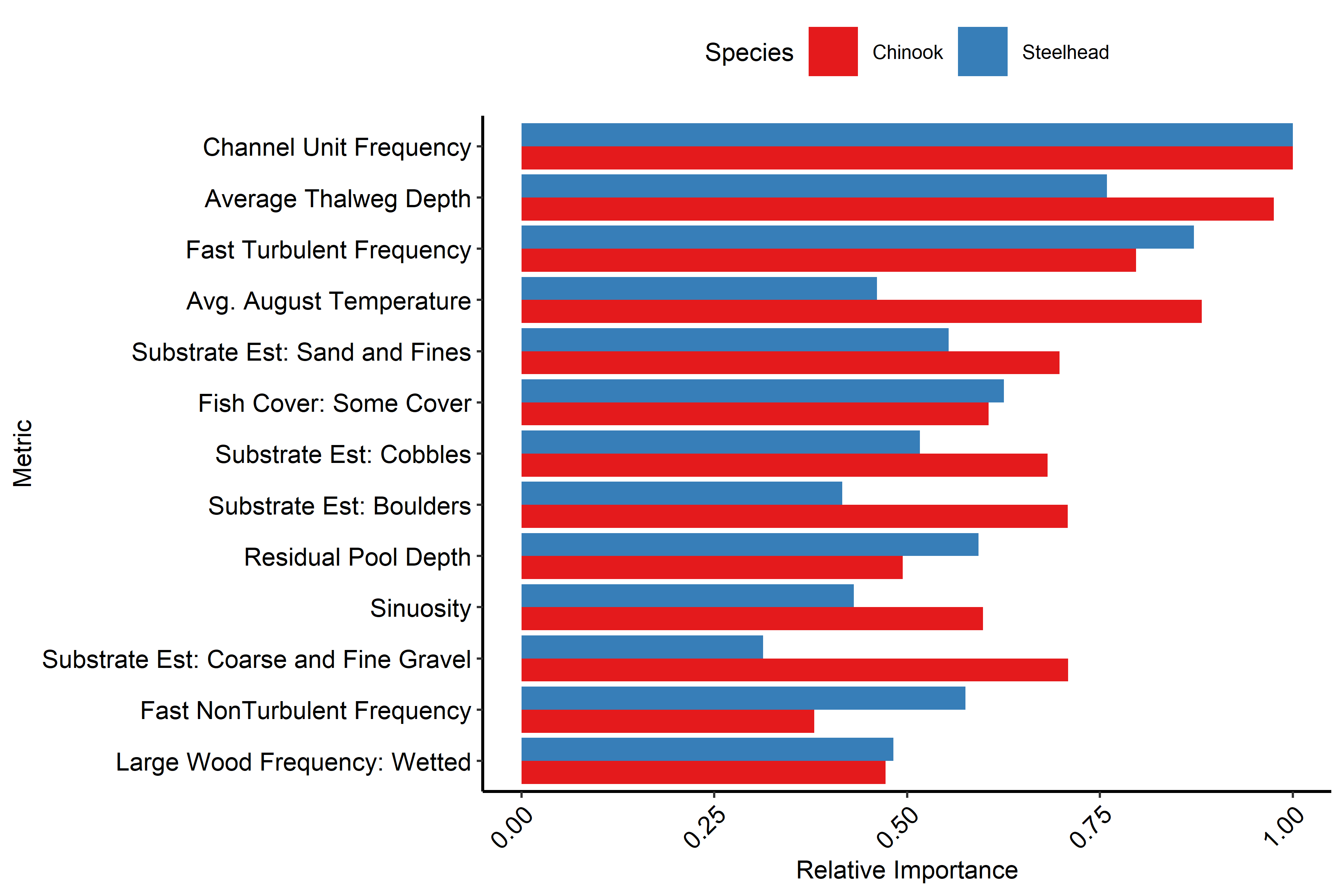


Figure 3: Relative importance plots for covariates included in the revised QRF redds models

Table 1: (#tab:cov-tab)Habitat covariates and their descriptions for three revised life stage QRF capacity models. Numbers indicate where each metric ranked in relative importance for each species. Dashes indicate the metric was not used for a given model.

Name

Metric Category

Juv Sum Chnk

Juv Sum Sthd

Juv Win Chnk

Juv Win Sthd

Redds Chnk

Redds Sthd

Description

Channel Unit Frequency

ChannelUnit

5

11

3

2

1

1

Number of channel units per 100 meters.

Fast NonTurbulent Frequency

ChannelUnit

9

13

–

–

13

6

Number of Fast Water Non-Turbulent channel units per 100 meters.

Fast Turbulent Frequency

ChannelUnit

3

6

–

–

4

2

Number of Fast Water Turbulent channel units per 100 meters.

Sinuosity

Complexity

13

7

6

5

10

11

Ratio of the thalweg length to the straight line distance between the start and end points of the thalweg.

Wetted Channel Braidedness

Complexity

14

14

10

11

–

–

Ratio of the total length of the wetted mainstem channel plus side channels and the length of the mainstem channel.

Fish Cover: LW

Cover

–

–

4

6

–

–

Percent of wetted area that has woody debris as fish cover.

Fish Cover: Some Cover

Cover

7

3

11

8

9

4

Percent of wetted area with some form of fish cover

Residual Depth

Size

–

–

2

3

–

–

Average residual depth of the channel unit.

Average Thalweg Depth

Size

1

2

–

–

2

3

Average Thalweg Depth, meters

Thalweg Exit Depth Avg

Size

–

–

5

4

–

–

Depth of the thalweg at the downstream edge of the channel unit.

Residual Pool Depth

Size

12

10

–

–

11

5

The average difference between the maximum depth and downstream end depth of all Slow Water/Pool channel units.

Discharge

Size

–

–

1

1

–

–

The sum of station discharge across all stations. Station discharge is calculated as depth x velocity x station increment for all stations except first and last. Station discharge for first and last station is 0.5 x station width x depth x velocity.

Substrate Est: Boulders

Substrate

8

9

–

–

6

12

Percent of boulders (256-4000 mm) within the wetted site area.

Substrate Est: Cobble and Boulder

Substrate

–

–

7

10

–

–

Total cobble plus boulder percentage

Substrate Est: Cobbles

Substrate

11

5

–

–

8

8

Percent of cobbles (64-256 mm) within the wetted site area.

Substrate Est: Coarse and Fine Gravel

Substrate

6

8

8

9

5

13

Percent of coarse and fine gravel (2-64 mm) within the wetted site area.

Substrate Est: Sand and Fines

Substrate

10

4

9

7

7

7

Percent of sand and fine sediment (0.01-2 mm) within the wetted site area.

Avg. August Temperature

Temperature

2

1

–

–

3

10

Average predicted daily August temperature from NorWest, averaged across the years 2002-2011.

Large Wood Frequency: Wetted

Wood

4

12

–

–

12

9

Number of large wood pieces per 100 meters within the wetted channel.

# Revised RF Extrapolation Model

The spatial extent of QRF capacity predictions is limited to reaches with high-resolution habitat data (i.e. CHaMP or DASH data). To estimate capacity outside of the QRF habitat capacity spatial extent, an extrapolation model fit to “globally available attributes” (GAAs) obtained from a continuous, linear stream network created by Morgan Bond and Tyler Nodine (<https://www.fisheries.noaa.gov/resource/data/columbia-basin-historical-ecology-project-data>) was used for the entire Columbia River Basin. A random forest model was fit using the GAAs from the linear stream network and used to estimate habitat capacity for the entire Columbia River Basin at a 200 meter reach scale. Consistent with the QRF habitat capacity models, the RF extrapolation model makes no assumptions about the direction and distribution of effects of predictors, and constrains capacity estimates within the range of predictions produced by the QRF habitat capacity model. However, random forest methods do not account for variable strata weights across the CHaMP dataset, a source of potential bias that could be alleviated through the collection of additional paired fish and habitat data.

RF extrapolation model covariates were selected from the list of GAAs and evaluated for inclusion by examining relative importance plots (Figure 4, Figure 5, and Figure 6 ), partial dependence plots (Section ?? ), and correlations between covariates. We used the previous extrapolation model as a starting point for covariate selection. CONTINUE DISCUSSION

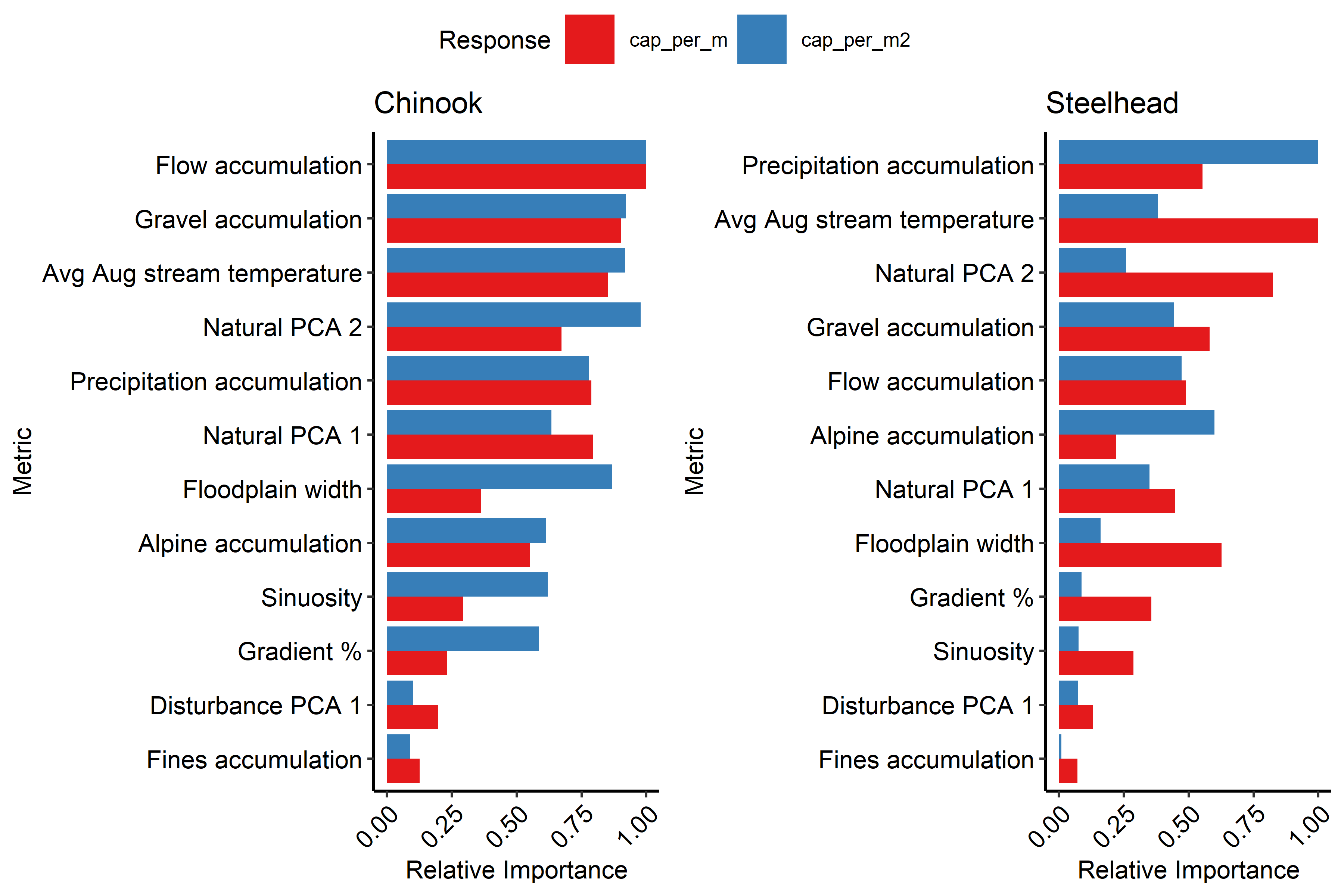


Figure 4: Relative importance plots for covariates included in the revised juvenile summer RF extrapolation models

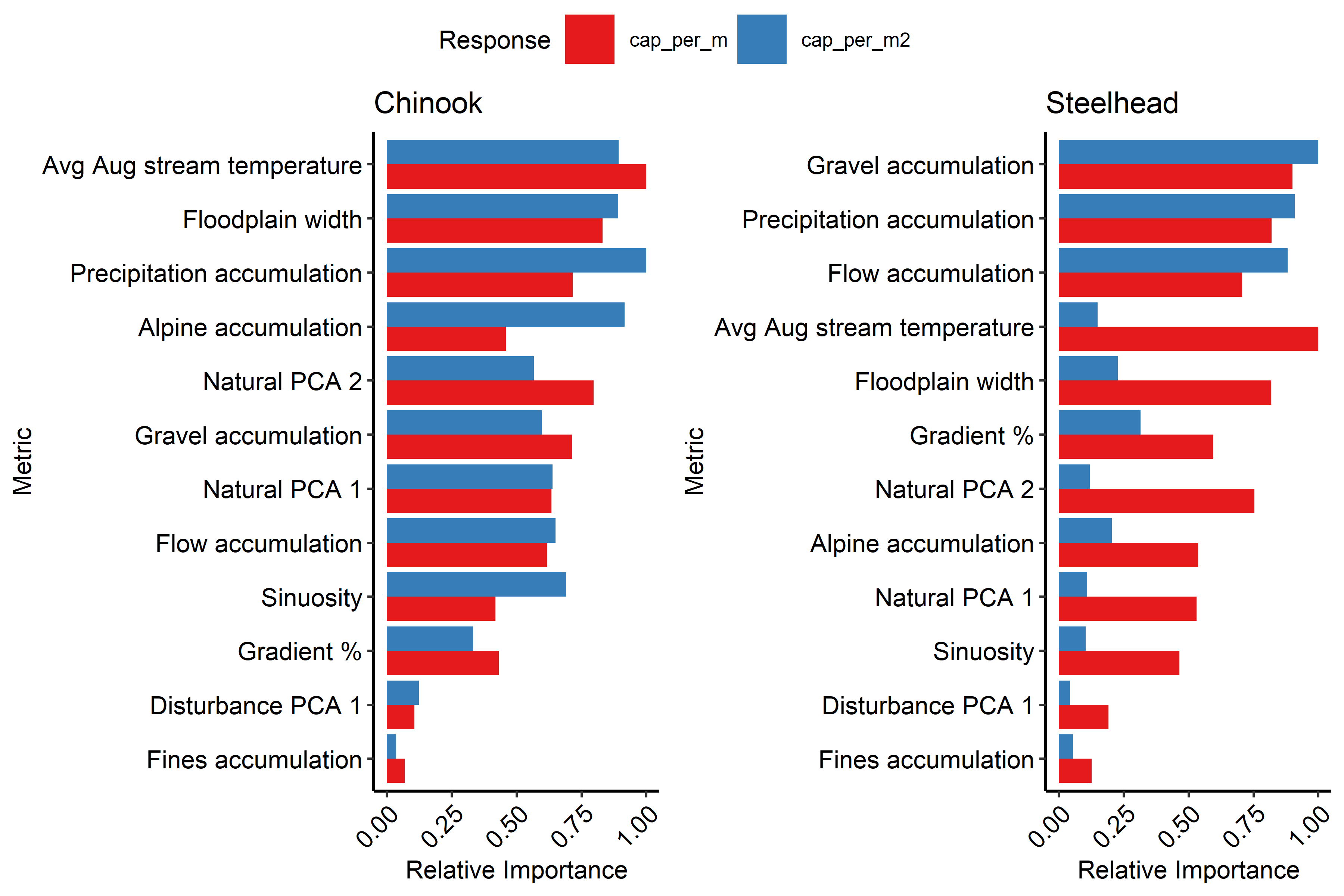


Figure 5: Relative importance plots for covariates included in the revised juvenile winter RF extrapolation models

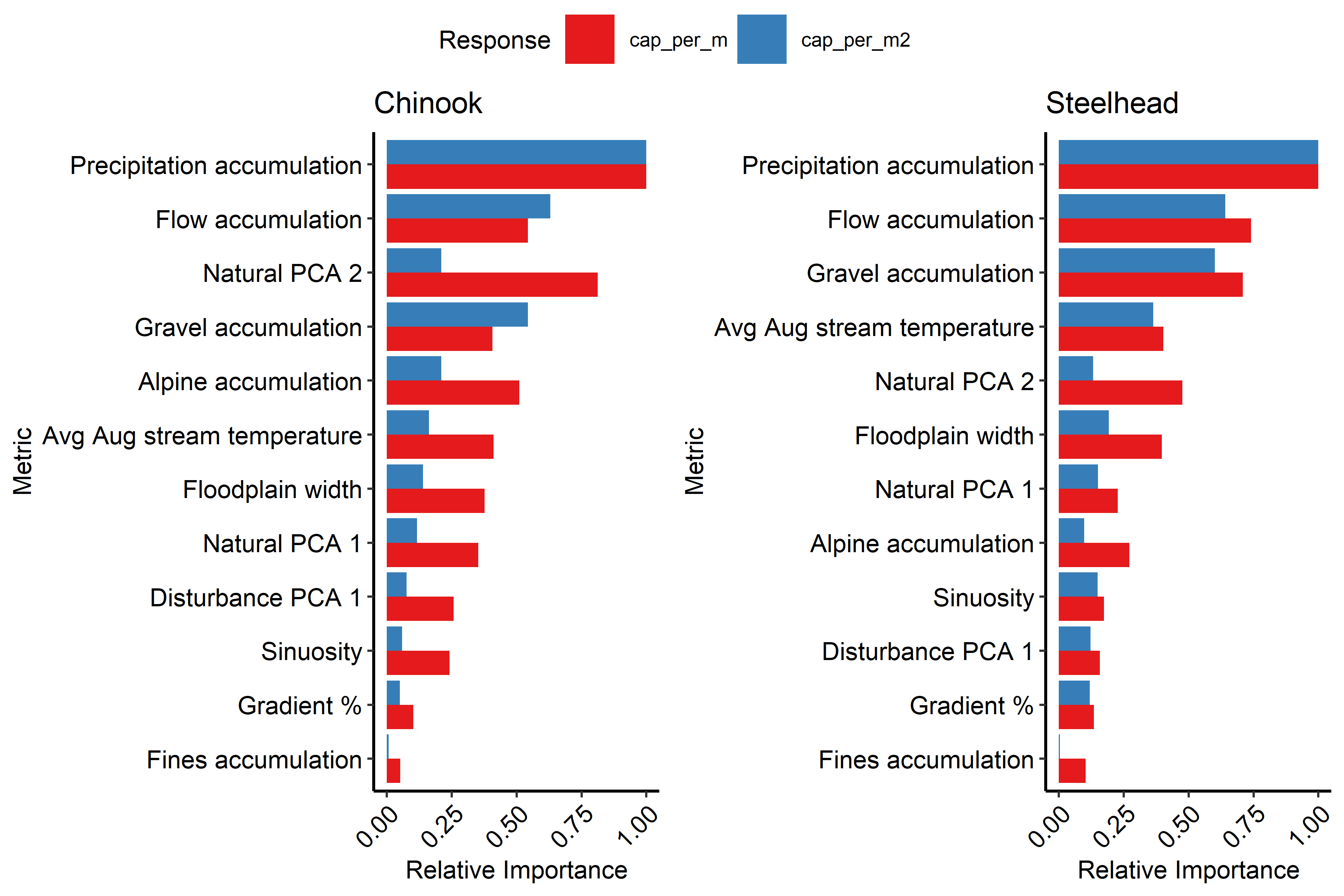


Figure 6: Relative importance plots for covariates included in the revised juvenile winter RF extrapolation models

Table 3: (#tab:gaa-tbl)Globally available attritibutes (GAAs) and their descriptions used in the random forest extrapolation model.

Metric

Decription

Gradient %

Stream gradient (%).

Sinuosity

Reach sinuosity. 1 = straight, 1 < sinuous.

Alpine accumulation

Number of upstream cells in alpine terrain.

Fines accumulation

Number of upstream cells in fine grain lithologies.

Flow accumulation

Number of upstream DEM cells flowing into reach.

Gravel accumulation

Number of upstream cells in gravel producing lithologies.

Precipitation accumulation

Number of upstream cells weighted by average annual precipitation.

Floodplain width

Current unmodified floodplain width.

Avg Aug stream temperature

Historical composite scenario representing 10 year average August mean stream temperatures for 2002-2011 (Isaak et al. 2017).

Disturbance PCA 1

Disturbance Classification PCA 1 Score (Whittier et al. 2011).

Natural PCA 1

Natural Classification PCA 1 Score (Whittier et al. 2011).

Natural PCA 2

Natural Classification PCA 2 Score (Whittier et al. 2011).

# Habitat capacity estimates

Habitat carrying capacity was estimated with the revised QRF and RF extrapolation models for Chinook salmon and steelhead during juvenile summer, juvenile winter, and redd life stages for eight watersheds in the Upper Salmon River Basin. Spatial domains for species were originally defined by Streamnet (<https://www.streamnet.org/home/data-maps/gis-data-sets/>) and spatial extents were revised based on expert knowledge from regional biologists.

## Chinook

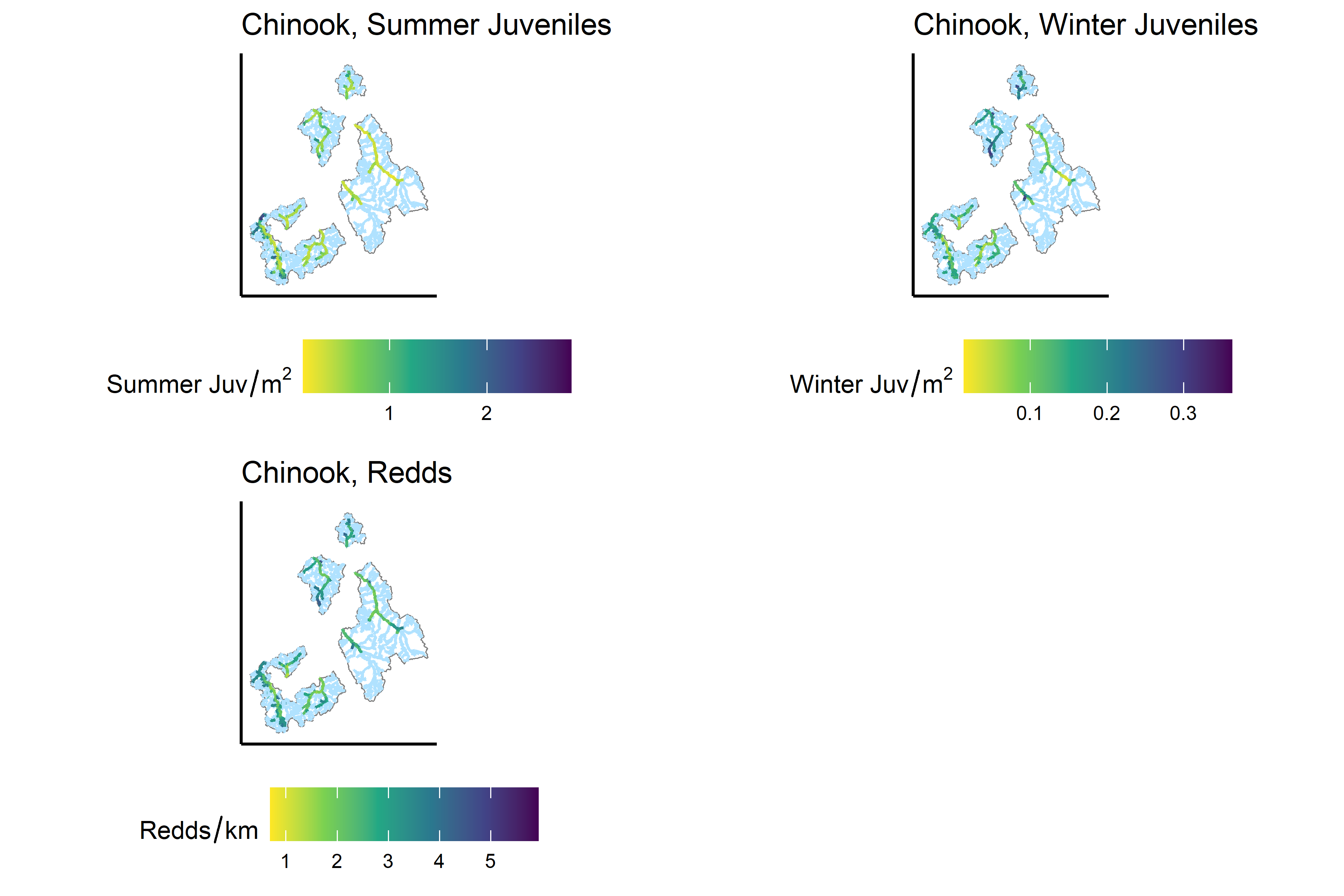


Figure 7: Extrapolations of habitat capacity for Chinook salmon, by life-stage, for the eight watersheds within the Upper Salmon River Basin using the revised models.

Table 5: (#tab:chnk-cap-summarytbl)Predicted Chinook salmon habitat capacity by life-stage and watershed using the revised models.

Watershed

Juv summer capacity

Summer SE

Juv winter capacity

Winter SE

Redd capacity

Redd SE

EF Salmon

623,619

75,423.7

117,246

17,703.8

252

11.0

Lemhi

580,216

36,467.0

144,888

19,358.4

314

11.0

NF Salmon

266,184

37,608.7

68,193

10,023.0

153

7.1

Pahsimeroi

333,212

31,459.0

129,357

18,763.4

184

6.8

Panther Cr

710,195

80,292.9

174,968

17,815.2

336

12.7

Upper Salmon

682,131

117,275.6

159,188

38,570.2

364

19.5

Valley Cr

562,444

91,839.2

113,782

29,705.4

312

15.9

Yankee Fork

245,014

34,108.2

50,003

8,878.2

129

5.3

Table 5: (#tab:chnk-cap-summarytbl)Predicted Chinook salmon habitat capacity per kilometer by life-stage and watershed using the revised models.

Watershed

Juv summer capacity/km

Summer SE/km

Juv winter capacity/km

Winter SE/km

Redd capacity/km

Redd SE/km

EF Salmon

5,798

701.2

1,090

164.6

2

0.1

Lemhi

4,416

277.6

1,103

147.3

2

0.1

NF Salmon

5,297

748.4

1,357

199.5

3

0.1

Pahsimeroi

4,701

443.8

1,825

264.7

3

0.1

Panther Cr

6,335

716.2

1,561

158.9

3

0.1

Upper Salmon

5,193

892.9

1,212

293.6

3

0.1

Valley Cr

5,839

953.4

1,181

308.4

3

0.2

Yankee Fork

4,581

637.7

935

166.0

2

0.1

## Steelhead

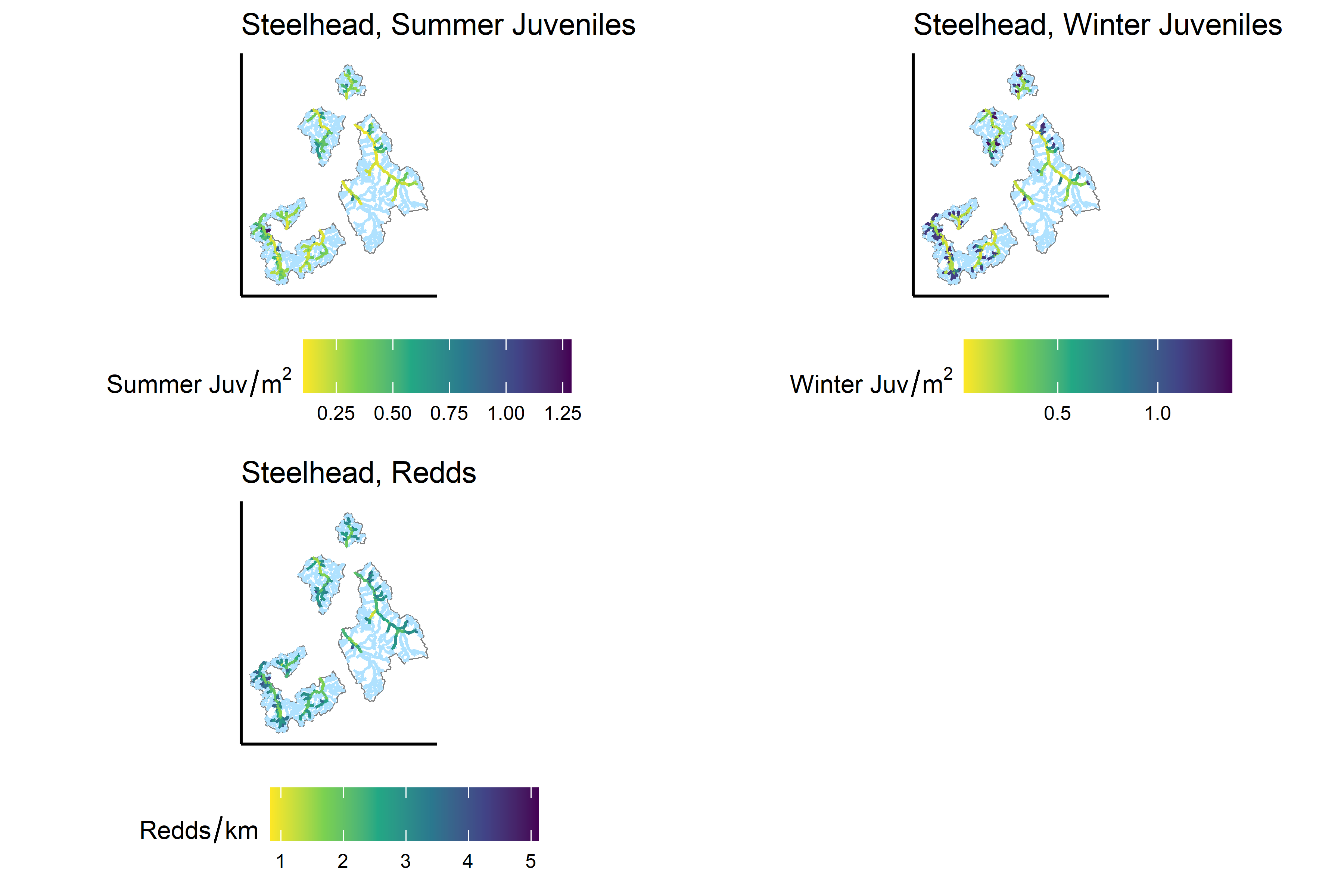


Figure 8: Extrapolations of habitat capacity for steelhead, by life-stage, for the eight watersheds within the Upper Salmon River Basin using the revised models.

Table 7: (#tab:sthd-cap-summarytbl)Predicted steelhead habitat capacity by life-stage and watershed using the revised models.

Watershed

Juv summer capacity

Summer SE

Juv winter capacity

Winter SE

Redd capacity

Redd SE

EF Salmon

238,197

16,910.5

285,190

27,988

336

17

Lemhi

443,662

16,517.1

586,570

53,623

667

30

NF Salmon

166,131

14,894.3

203,030

21,468

212

14

Pahsimeroi

157,397

5,851.4

187,196

13,663

163

8

Panther Cr

273,969

19,130.0

385,158

35,799

362

23

Upper Salmon

258,297

16,489.4

327,831

38,972

440

30

Valley Cr

184,939

11,975.6

275,905

28,991

348

24

Yankee Fork

112,944

8,675.2

133,109

17,008

173

11

Table 7: (#tab:sthd-cap-summarytbl)Predicted steelhead habitat capacity per kilometer by life-stage and watershed using the revised models.

Watershed

Juv summer capacity/km

Summer SE/km

Juv winter capacity/km

Winter SE/km

Redd capacity/km

Redd SE/km

EF Salmon

1,732

123.0

2,074

203.5

2

0.1

Lemhi

1,754

65.3

2,319

212.0

3

0.1

NF Salmon

2,046

183.4

2,500

264.4

3

0.2

Pahsimeroi

2,233

83.0

2,656

193.8

2

0.1

Panther Cr

1,920

134.1

2,700

250.9

3

0.2

Upper Salmon

1,634

104.3

2,074

246.6

3

0.2

Valley Cr

1,649

106.8

2,461

258.5

3

0.2

Yankee Fork

1,658

127.4

1,954

249.7

3

0.2

# Habitat capacity estimates compared with previous QRF and extrapolation

Comparisons of watershed capacity estimates between the previous and revised QRF and RF extrapolation models reveal modest differences in most cases, with the exception of Chinook parr summer capacities in several watersheds. More discussion here…

## Chinook

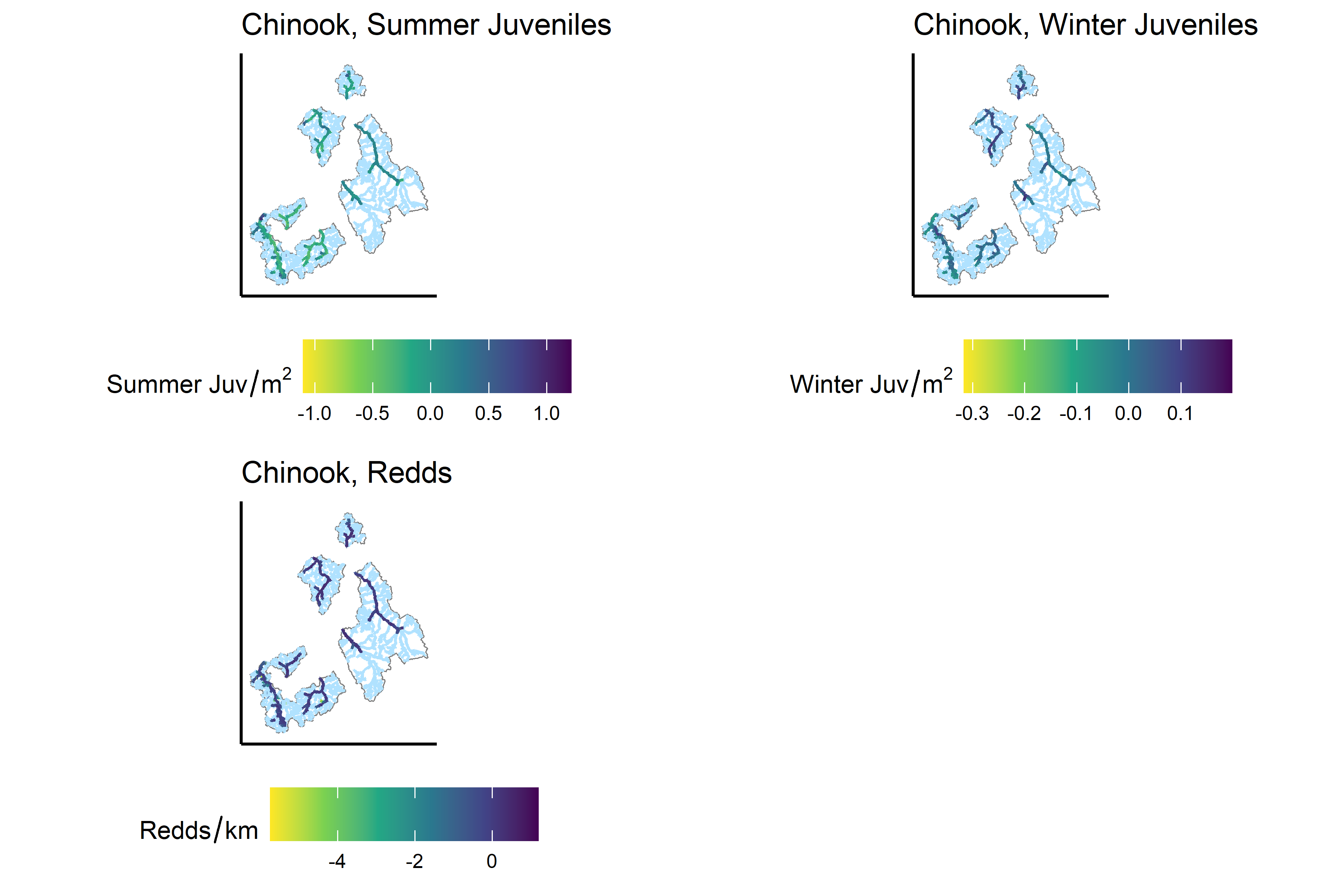


Figure 9: Change in predicted Chinook salmon habitat capacity estimates from the original model and extrapolation, by life-stage, for the eight watersheds within the Upper Salmon River Basin.

Table 9: (#tab:chnk-comp-summary-tbls)Estimated Chinook salmon capacities and comparison with previous random forest extrapolations for eight watersheds

Model

Watershed

Capacity per km

Total capacity

Capacity % change

Capacity SE

Juv summer

EF Salmon

5,797.6

623,619

-10

75,424

Lemhi

4,416.3

580,216

64

36,467

NF Salmon

5,297.3

266,184

-14

37,609

Pahsimeroi

4,700.8

333,212

24

31,459

Panther Cr

6,335.1

710,195

-14

80,293

Upper Salmon

5,193.3

682,131

-24

117,276

Valley Cr

5,838.6

562,444

-12

91,839

Yankee Fork

4,581.0

245,014

-34

34,108

Juv winter

EF Salmon

1,090.0

117,246

20

17,704

Lemhi

1,102.8

144,888

-3

19,358

NF Salmon

1,357.1

68,193

25

10,023

Pahsimeroi

1,824.9

129,357

8

18,763

Panther Cr

1,560.8

174,968

40

17,815

Upper Salmon

1,211.9

159,188

3

38,570

Valley Cr

1,181.1

113,782

4

29,705

Yankee Fork

934.9

50,003

14

8,878

Redds

EF Salmon

2.3

252

-8

11

Lemhi

2.4

314

-2

11

NF Salmon

3.0

153

1

7

Pahsimeroi

2.6

184

5

7

Panther Cr

3.0

336

3

13

Upper Salmon

2.8

364

-11

19

Valley Cr

3.2

312

-21

16

Yankee Fork

2.4

129

-5

5

## Steelhead

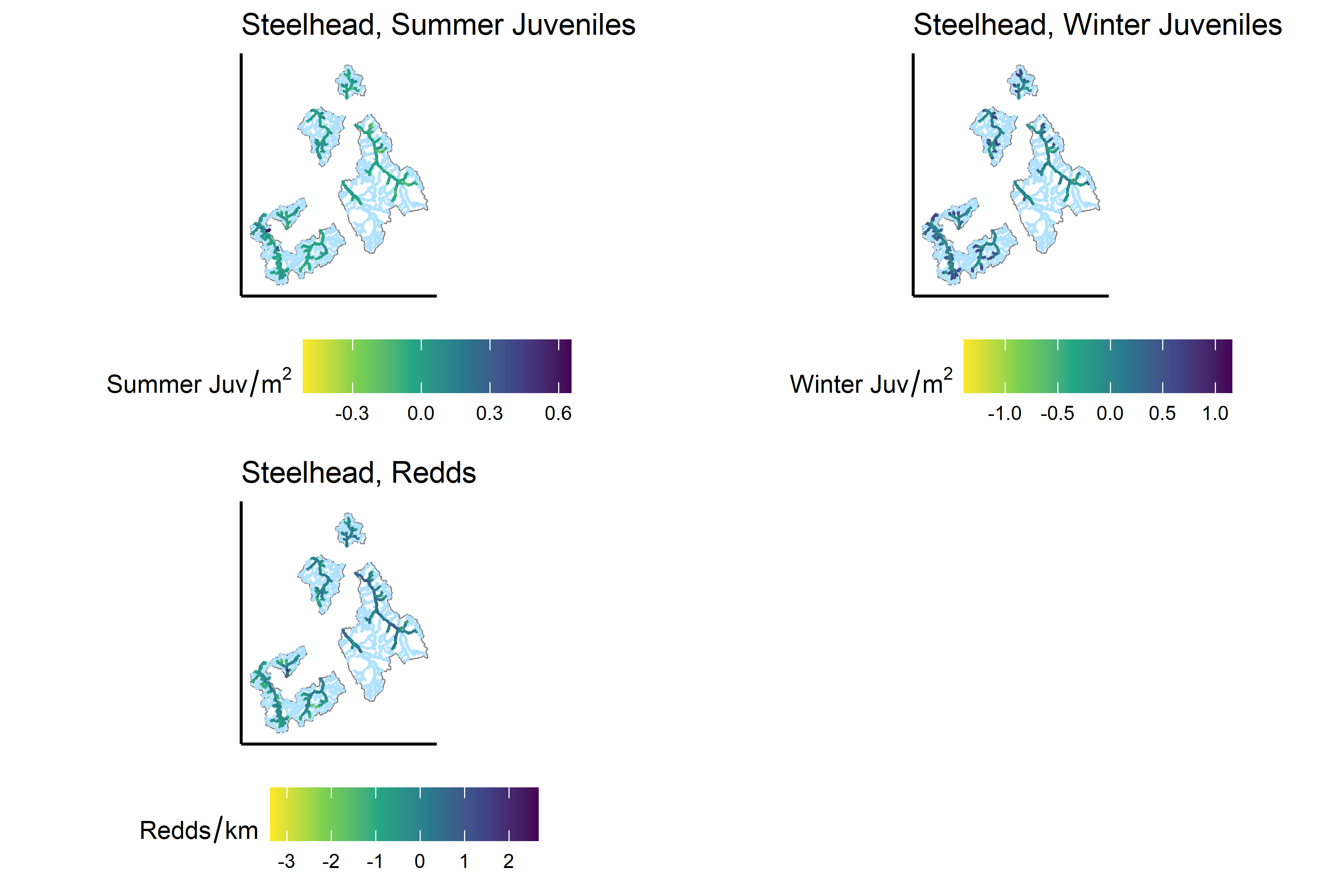


Figure 10: Change in steelhead habitat capacity estimates from the original model and extrapolation, by life-stage, for the eight watersheds within the Upper Salmon River Basin.

Table 11: (#tab:sthd-comp-summary-tbls)Estimated steelhead capacities and comparison with previous random forest extrapolations for eight watersheds

Model

Watershed

Capacity per km

Total capacity

Capacity % change

Capacity SE

Juv summer

EF Salmon

1,732.2

238,197

-22

16,911

Lemhi

1,754.0

443,662

-19

16,517

NF Salmon

2,046.0

166,131

-18

14,894

Pahsimeroi

2,233.1

157,397

-10

5,851

Panther Cr

1,920.4

273,969

-17

19,130

Upper Salmon

1,634.4

258,297

-24

16,489

Valley Cr

1,649.3

184,939

-20

11,976

Yankee Fork

1,658.3

112,944

-24

8,675

Juv winter

EF Salmon

2,073.9

285,190

-10

27,988

Lemhi

2,319.0

586,570

-5

53,623

NF Salmon

2,500.4

203,030

1

21,468

Pahsimeroi

2,655.9

187,196

6

13,663

Panther Cr

2,699.8

385,158

11

35,799

Upper Salmon

2,074.4

327,831

-19

38,972

Valley Cr

2,460.5

275,905

-11

28,991

Yankee Fork

1,954.4

133,109

-10

17,008

Redds

EF Salmon

2.4

336

-11

17

Lemhi

2.6

667

8

30

NF Salmon

2.6

212

0

14

Pahsimeroi

2.3

163

15

8

Panther Cr

2.5

362

-4

23

Upper Salmon

2.8

440

-10

30

Valley Cr

3.1

348

-17

24

Yankee Fork

2.5

173

-10

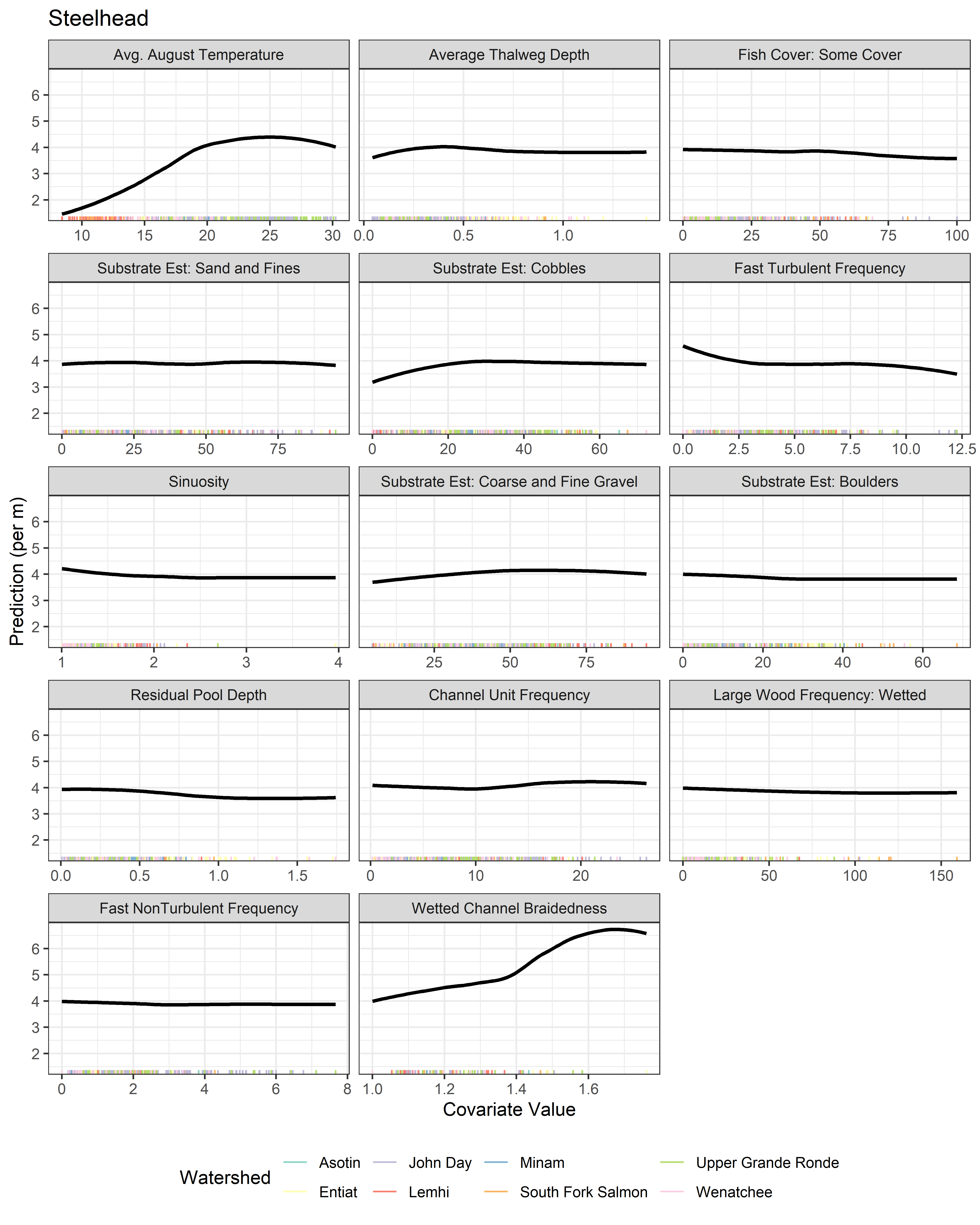
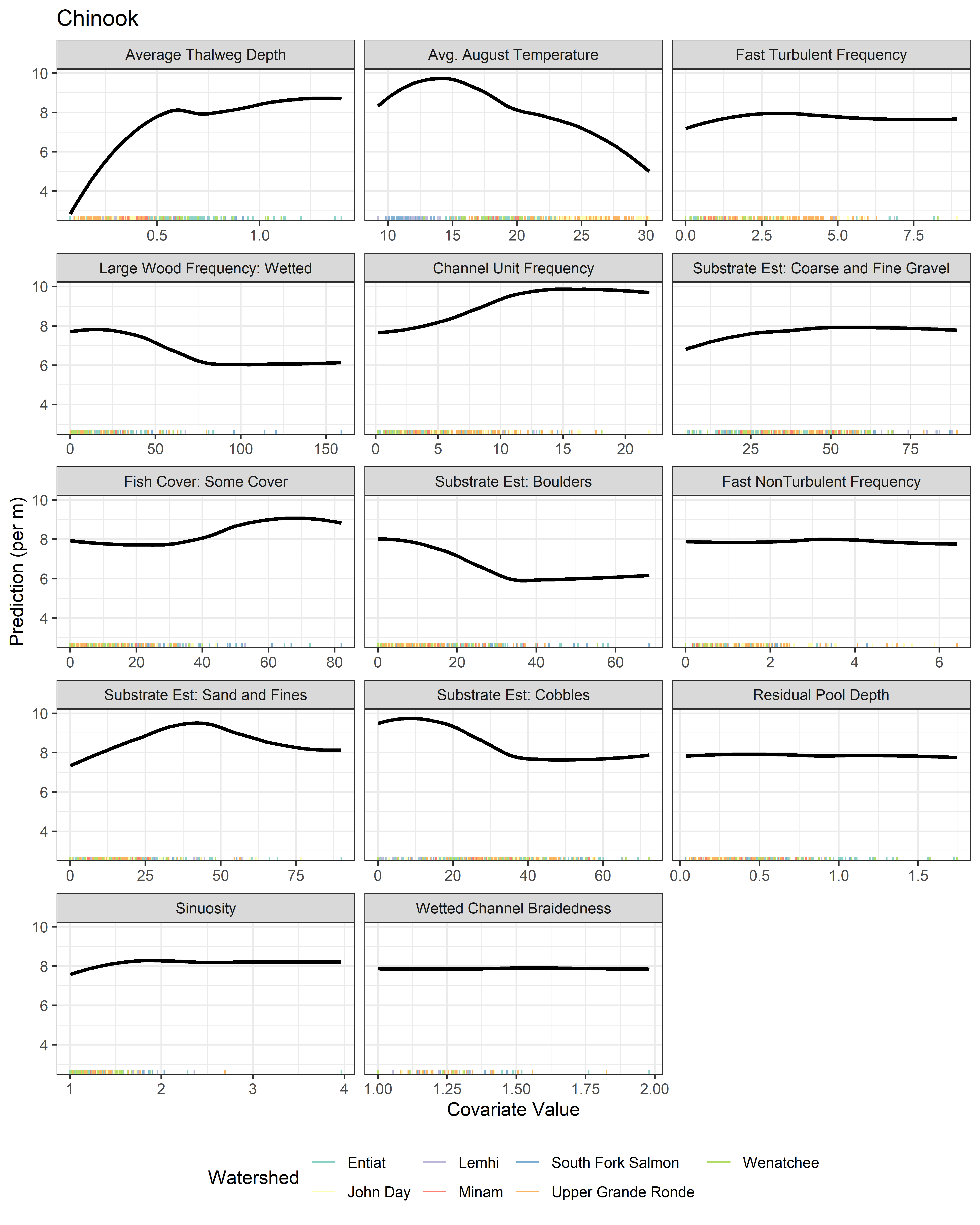
11

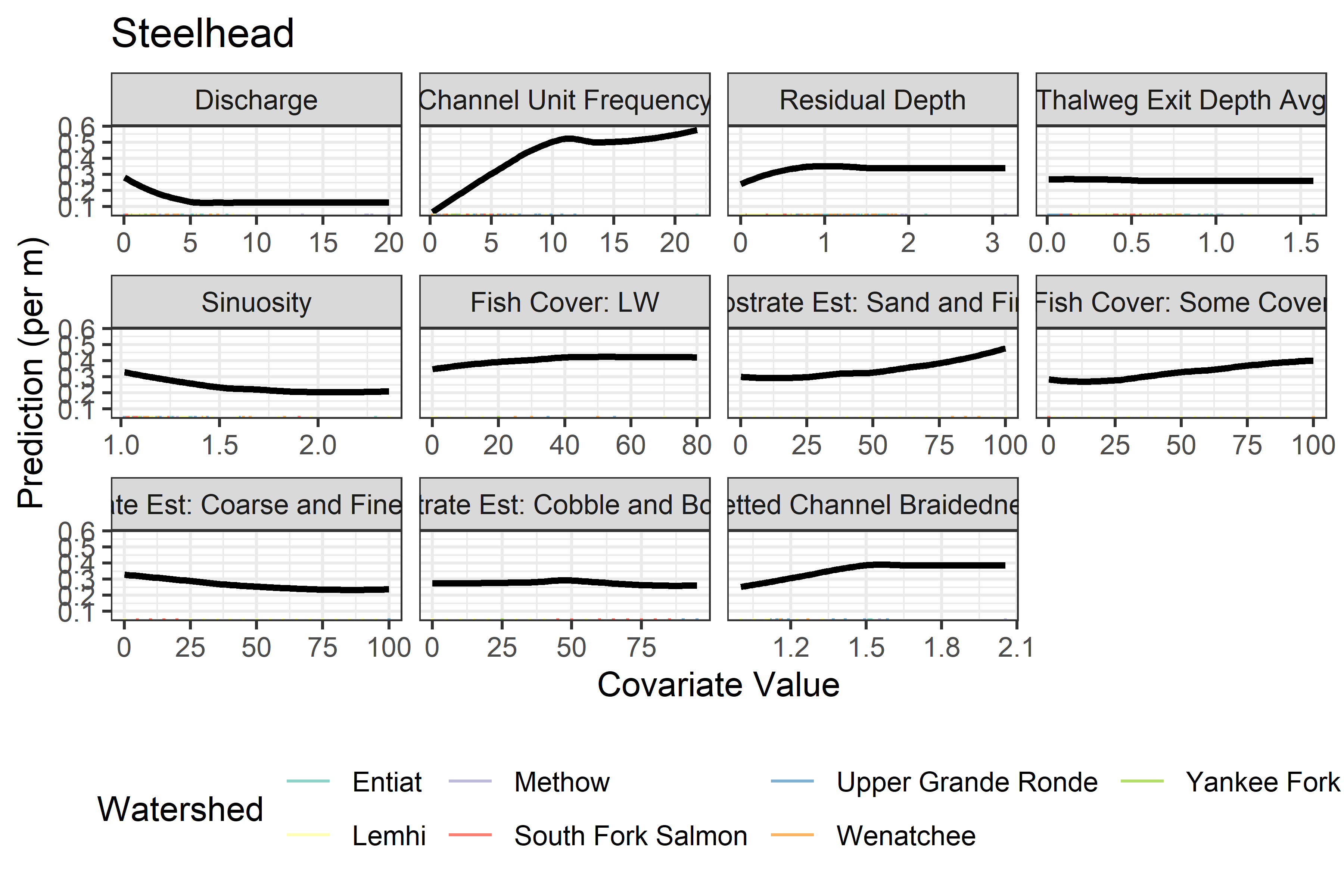
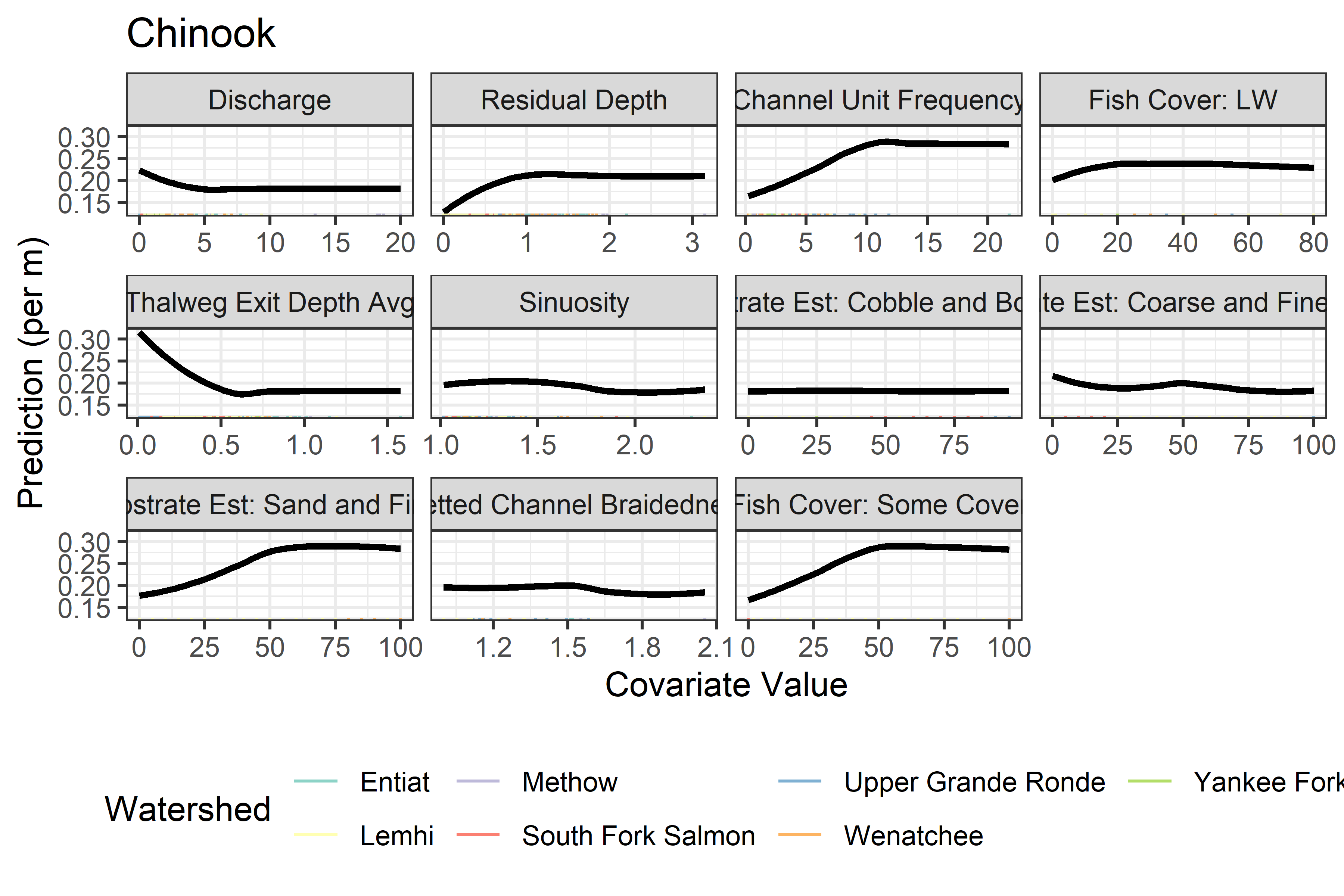
# Supplemental Figures and Tables

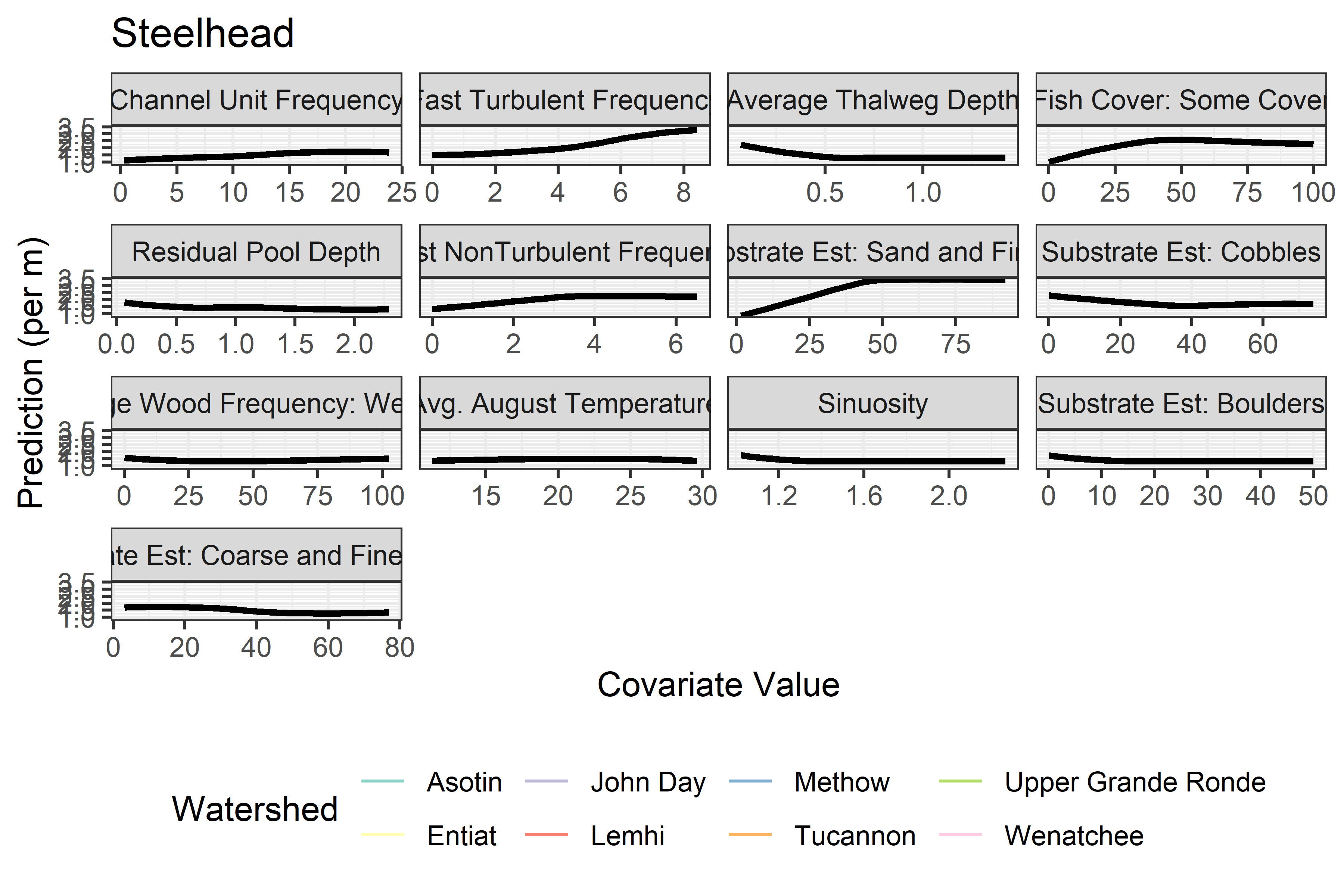
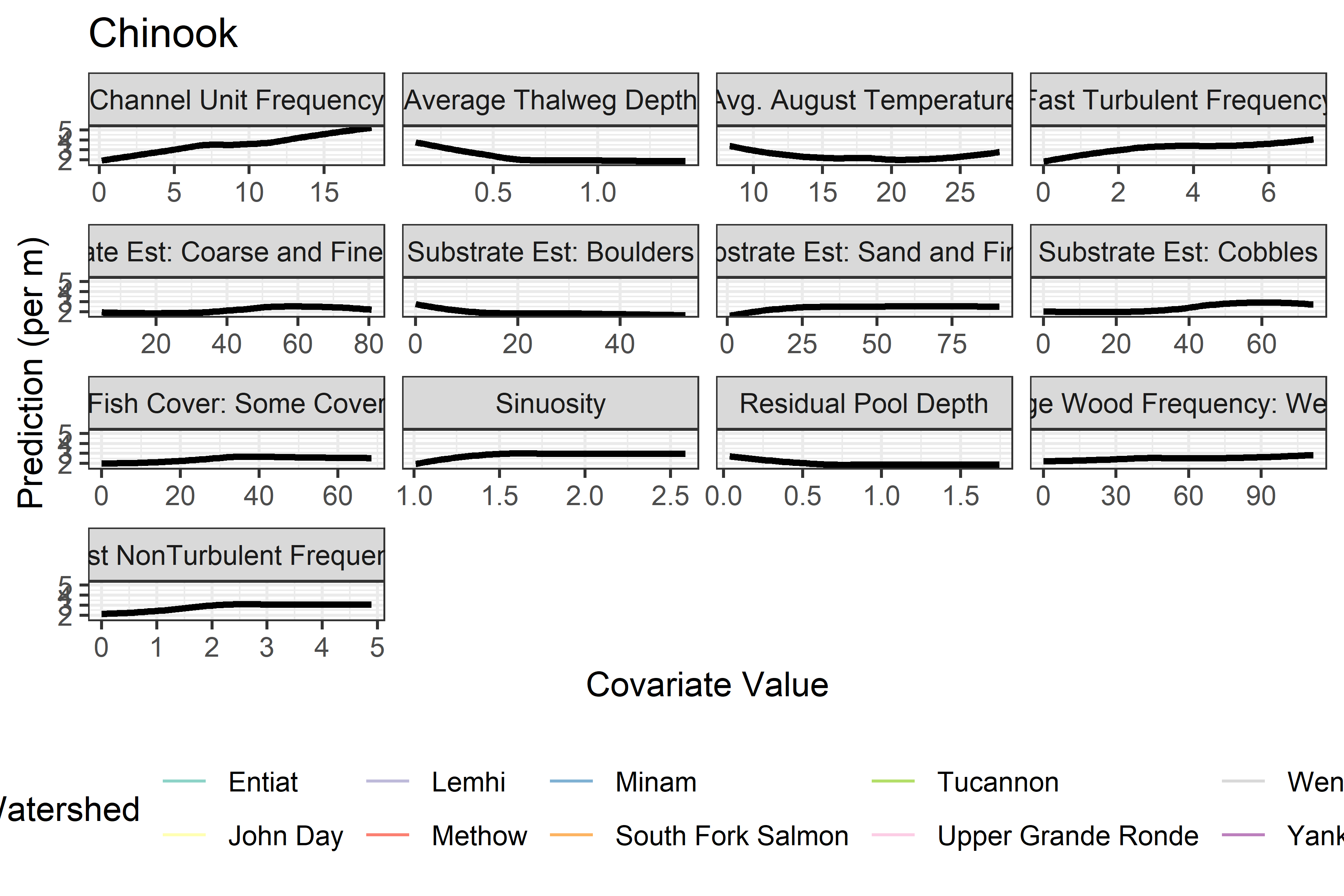
## Partial dependence plots

To support the covariate selection process for the QRF capacity and RF extrapolation models,partial dependence plots were generated to illustrate the predicted effect of covariates on fish density and capacity. Partial dependence plots function similarly to traditional covariate effects plots where predictions on the response are made by altering the value of the covariate of interest while all others are fixed at mean values. Because random forest models do not place any constraints on the possible mathematical relationships between predictor and response variables, effects curves have been visualized using smoothing methods (LOESS) and may not reflect actual model behavior across the range of covariate values.

### Revised QRF habitat capacity model







### Revised RF extrapolation model

