
Techniques to improve ecological interpretability
of black-box machine learning models

Supplemental material

Authors

[REDACTED]

Affiliations

[REDACTED]

[REDACTED]

[REDACTED]

September 29, 2021

1 Case study

Identifier	Description
Ecoregion	NARS 9-level reporting region (2015), based on aggregated Omernik Level III ecoregions
Benthic MMI condition	Condition class based on benthic macroinvertebrate score index (response)
Bed stability	Log relative bed stability
Riparian vegetation condition	Log sum of Woody Canopy + Mid + Ground layer areal cover proportion used for condition
Fish cover	Log sum of non-anthropogenic fish areal cover types used for condition
Elevation	Mean elevation value in meters (NED 30m resolution) within NHDPlus navigated cumulative basin
NO ₃ deposition	Mean value for Nitrate ion wet deposition, 2009 - NHDPlus cumulative watershed geometry
SO ₄ deposition	Mean value for Sulfate ion wet deposition, 2009 - NHDPlus cumulative watershed geometry
Tree canopy	Mean percent tree canopy (NLCD2001 tree canopy) within NHDPlus navigated cumulative basin
Impervious surface	Mean percent impervious surface (NLCD2006 Impervious Surface) within NHDPlus navigated cumulative basin
Sandy soils	Mean percent sand (US EPA Atlas of Ecosystem Services - SSURGO & STATSGO2) within NHDPlus navigated cumulative basin
Catchment slope	Mean percent slope (NED 30m resolution) within NHDPlus navigated cumulative basin
Agriculture	Percentage of NHDPlus navigated cumulative basin area classified as Pasture/Hay (81) or Row Crops (82) in NLCD
Shrub/Grass	Percentage of NHDPlus navigated cumulative basin area classified as Shrub/Scrub (52) or Grasslands/Herbaceous (71) in NLCD
Wetlands	Percentage of NHDPlus navigated cumulative basin area classified as Woody Wetlands (90), or Emergent Herbaceous Wetlands (95) in NLCD
Max temperature	Average maximum annual temperature in degrees Celsius (PRISM 1997-2000) at the sample site
Human Disturbance Index	Human Disturbance Index (distance-wtd tally of types and presence used for condition, PHAB)
Watershed area	NARS watershed area used in indicator analyses for NRSA 2008-9

Table S1: Variable description of all covariates of the case study

Identifier	Description
ALE	Accumulated local effects
EPA	Environmental Protection Agency
GBT	Gradient boosted trees
GAMs	Generalized additive models
GLMs	Generalized linear models
IAS	Interaction strength
ICE	Individual conditional expectation
MMI	Multimetric index
NRSA	National Rivers and Streams Assessment
PDP	partial dependence plots
USGS	U.S. Geological Survey

Table S2: Table of abbreviations used in the article.

1.1 Significant interaction effects

Variable1	Variable2	H: Benthic MMI condition=Fair	Adjusted p-value
NO ₃ deposition	SO ₄ deposition	0.4138	$< 10^{-4}$
Elevation	Max temperature	0.2965	$< 10^{-4}$
Catchment slope	Agriculture	0.2913	$< 10^{-4}$
Elevation	Catchment slope	0.2845	$< 10^{-4}$
Watershed area	Agriculture	0.2570	$< 10^{-4}$
Tree canopy	Max temperature	0.2411	$< 10^{-4}$
NO ₃ deposition	Catchment slope	0.1949	$< 10^{-4}$
Catchment slope	Impervious surface	0.1931	$< 10^{-4}$
Catchment slope	Max temperature	0.1857	$4 * 10^{-4}$
Riparian vegetation condition	Tree canopy	0.1801	$3 * 10^{-4}$
Elevation	Impervious surface	0.1680	10^{-3}
Riparian vegetation condition	Agriculture	0.1596	$3 * 10^{-4}$
Elevation	Watershed area	0.1549	$4 * 10^{-4}$
Bed stability	Agriculture	0.1493	$5 * 10^{-4}$
NO ₃ deposition	Max temperature	0.1428	$8 * 10^{-4}$
Bed stability	NO ₃ deposition	0.1356	$9 * 10^{-4}$
SO ₄ deposition	Catchment slope	0.1272	$7 * 10^{-4}$

Table S3: Two-way interactions influencing Fair condition

Variable1	Variable2	H: Benthic MMI condition=Good	Corrected p-value
Ecoregion	Watershed area	0.2592	$< 10^{-4}$
Bed stability	Watershed area	0.2152	$< 10^{-4}$

Table S4: Two-way interactions influencing Good condition

Variable1	Variable2	Variable3	H: Benthic MMI condition=Fair	Adjusted p-value
SO ₄ deposition	Elevation	Catchment slope	0.0364	$< 10^{-4}$
Catchment slope	Watershed area	Agriculture	0.0328	$< 10^{-4}$
Elevation	Catchment slope	Agriculture	0.0316	$< 10^{-4}$
Ecoregion	Elevation	Wetlands	0.0310	$< 10^{-4}$
NO ₃ deposition	Elevation	Catchment slope	0.0306	$< 10^{-4}$
Elevation	Catchment slope	Shrub/Grass	0.0271	$< 10^{-4}$
NO ₃ deposition	Catchment slope	Max temperature	0.0250	$< 10^{-4}$
Tree canopy	Max temperature	Agriculture	0.0233	$2 * 10^{-4}$
SO ₄ deposition	Elevation	Max temperature	0.0227	$2 * 10^{-4}$
Catchment slope	Max temperature	Shrub/Grass	0.0227	$< 10^{-4}$
Human Disturbance Index	Catchment slope	Agriculture	0.0226	$2 * 10^{-4}$
NO ₃ deposition	Catchment slope	Agriculture	0.0225	$< 10^{-4}$
Riparian vegetation condition	Tree canopy	Agriculture	0.0225	$< 10^{-4}$
Elevation	Catchment slope	Watershed area	0.0211	$2 * 10^{-4}$
Bed stability	Agriculture	Wetlands	0.0201	$< 10^{-4}$
Bed stability	Tree canopy	Agriculture	0.0197	$2 * 10^{-4}$
Riparian vegetation condition	Tree canopy	Wetlands	0.0197	$< 10^{-4}$
Riparian vegetation condition	Tree canopy	Max temperature	0.0196	$2 * 10^{-4}$
Catchment slope	Agriculture	Wetlands	0.0194	$2 * 10^{-4}$
Riparian vegetation condition	Human Disturbance Index	Agriculture	0.0191	$2 * 10^{-4}$
NO ₃ deposition	Max temperature	Agriculture	0.0191	$2 * 10^{-4}$
Catchment slope	Tree canopy	Agriculture	0.0187	$2 * 10^{-4}$
Elevation	Tree canopy	Max temperature	0.0160	$2 * 10^{-4}$

Table S5: Three-way interactions influencing Fair condition

1.2 Additional ICE plots

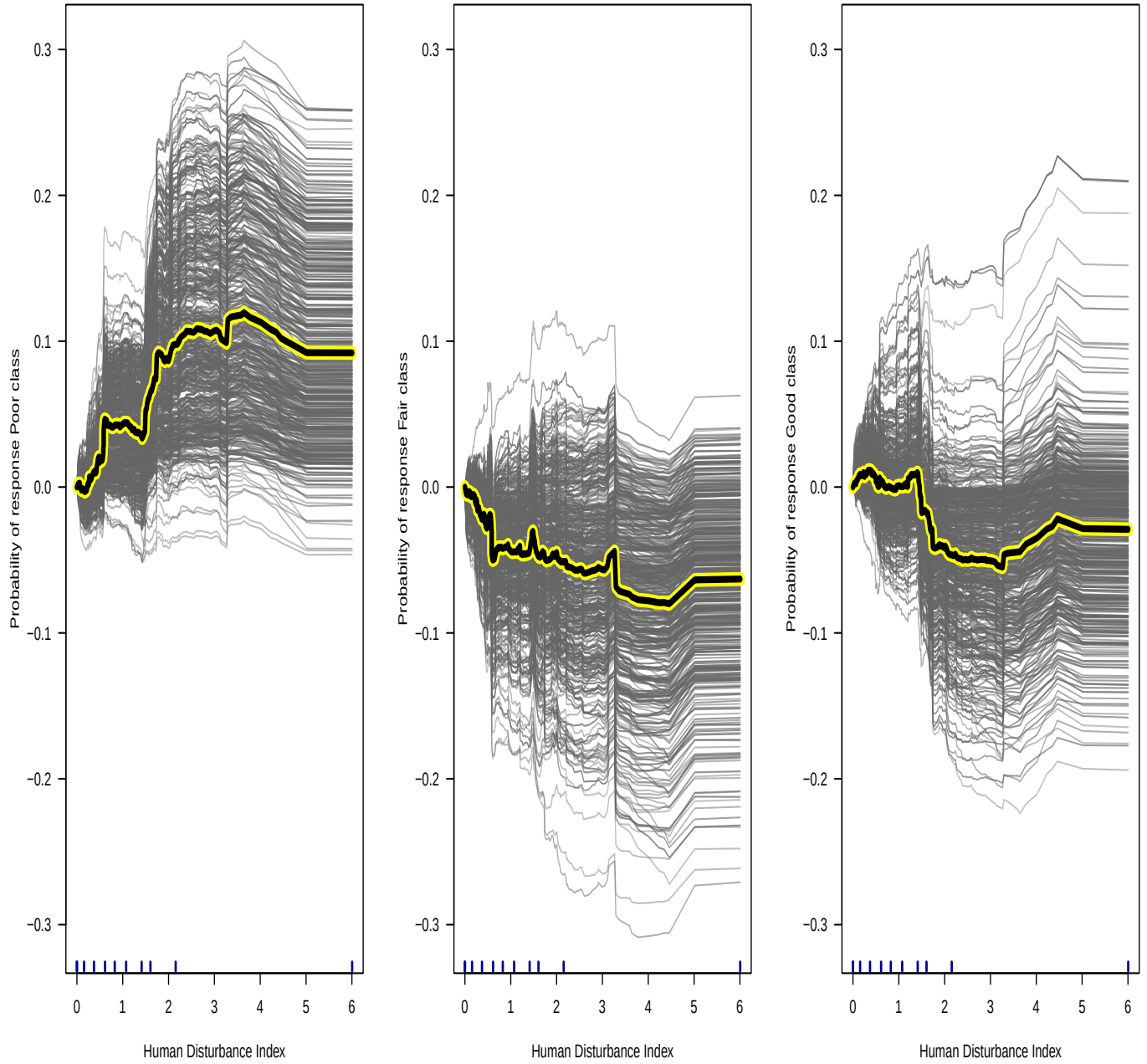


Figure S1: Centered marginal effect of human disturbance index on estimated benthic MMI condition distribution. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves).

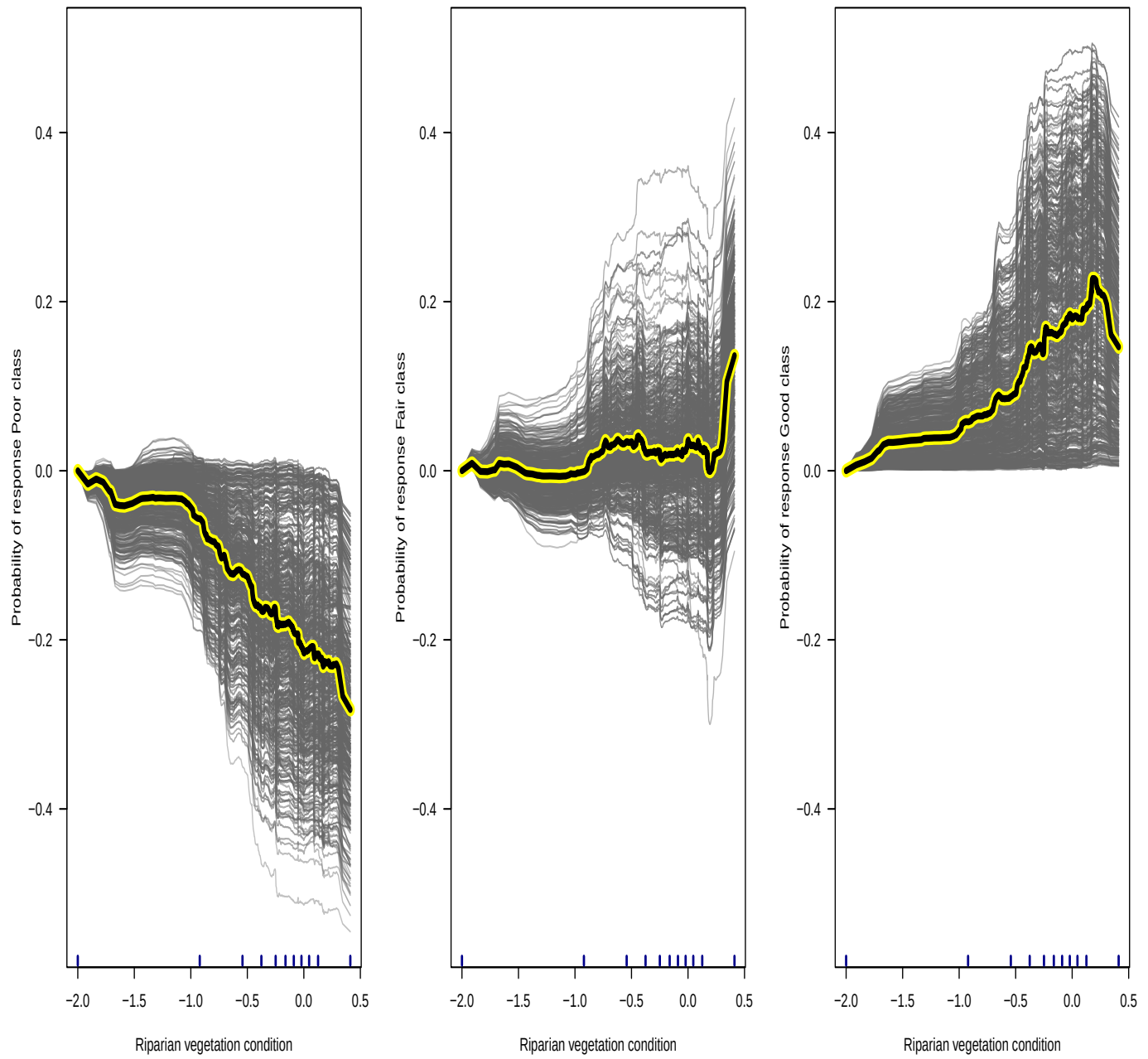


Figure S2: Centered marginal effect of riparian vegetation condition estimated benthic MMI condition distribution. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves).

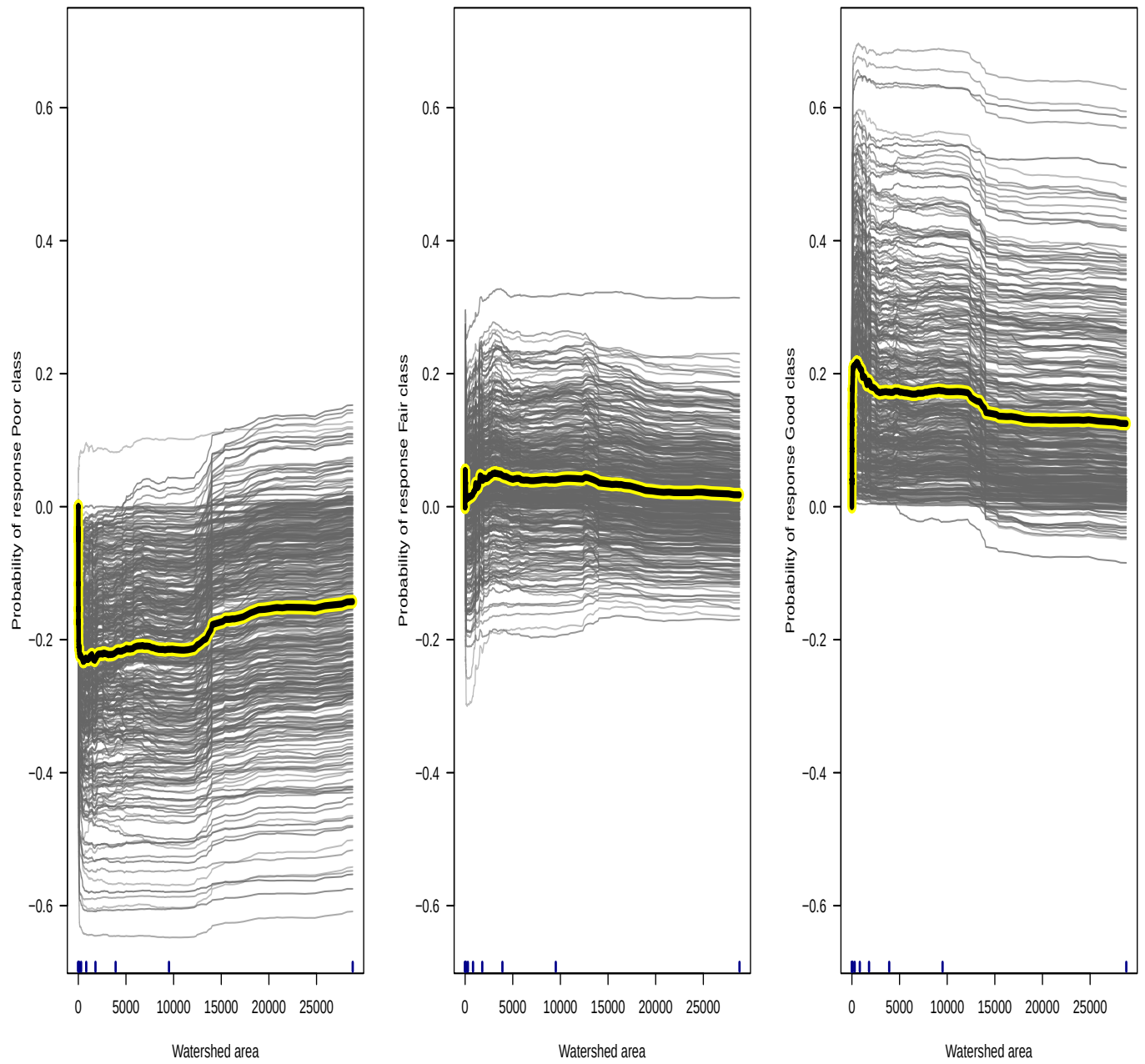


Figure S3: Centered marginal effect of watershed area on estimated benthic MMI condition distribution. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves).

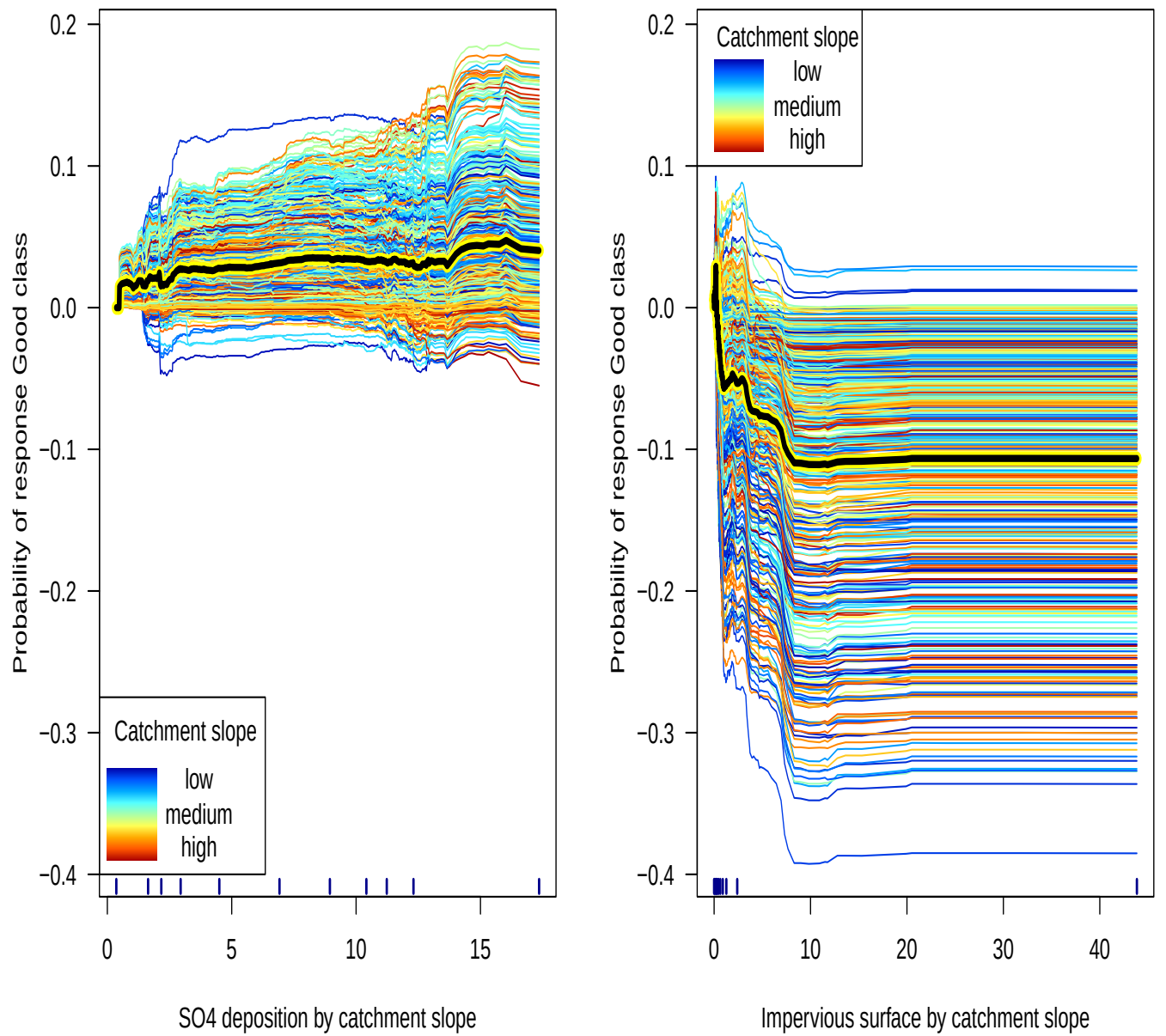


Figure S4: Centered marginal effect of impervious surface (right) on estimated benthic MMI condition distribution "Good". The colors represent low values (blue) to high values (red) of the second variable. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves).

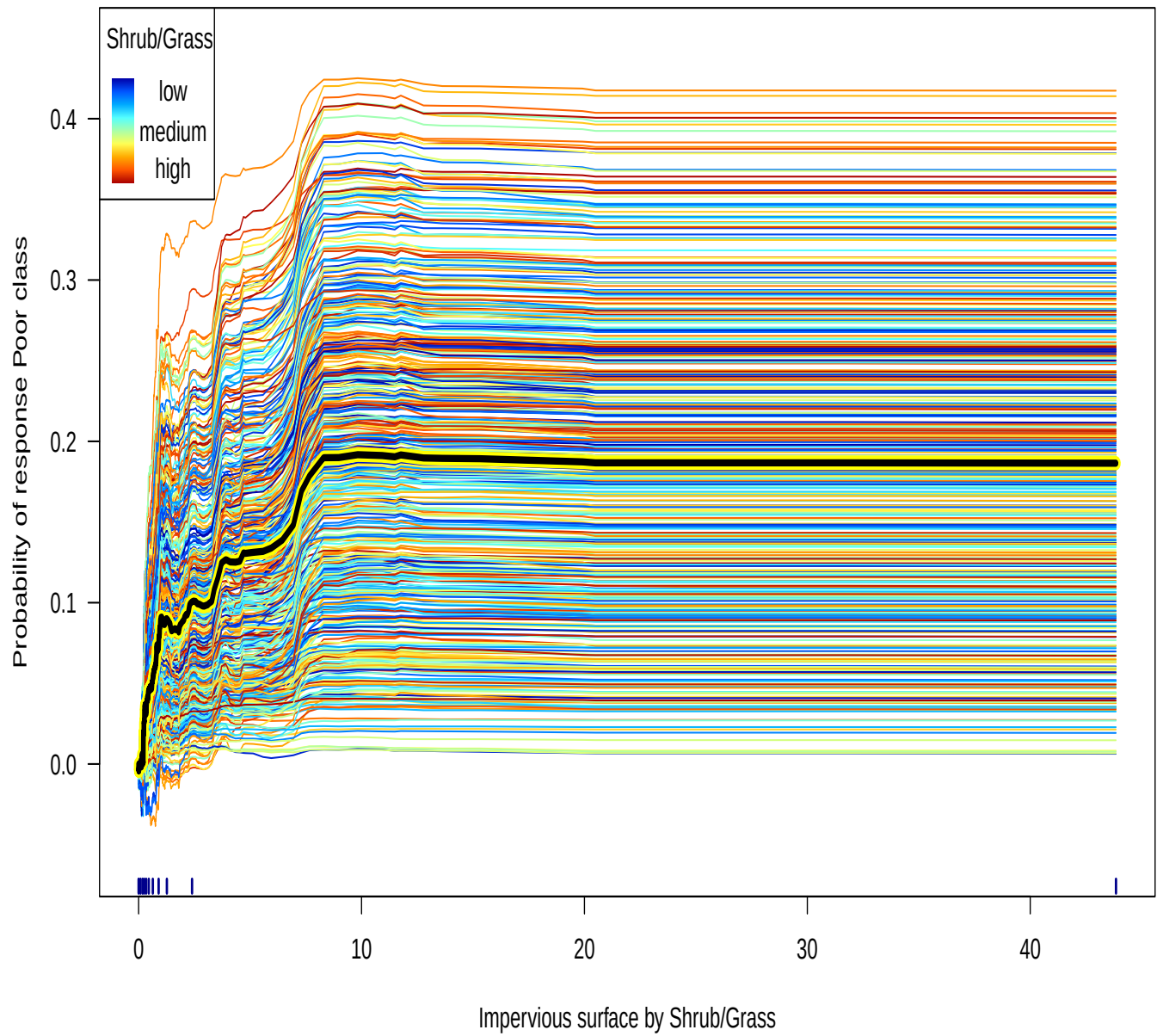


Figure S5: Centered marginal effect of impervious surface (right) on estimated benthic MMI condition distribution "Poor". The colors represent low values (blue) to high values (red) of the second variable. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves).

1.3 ICE plots with probability integral transformed x-axis

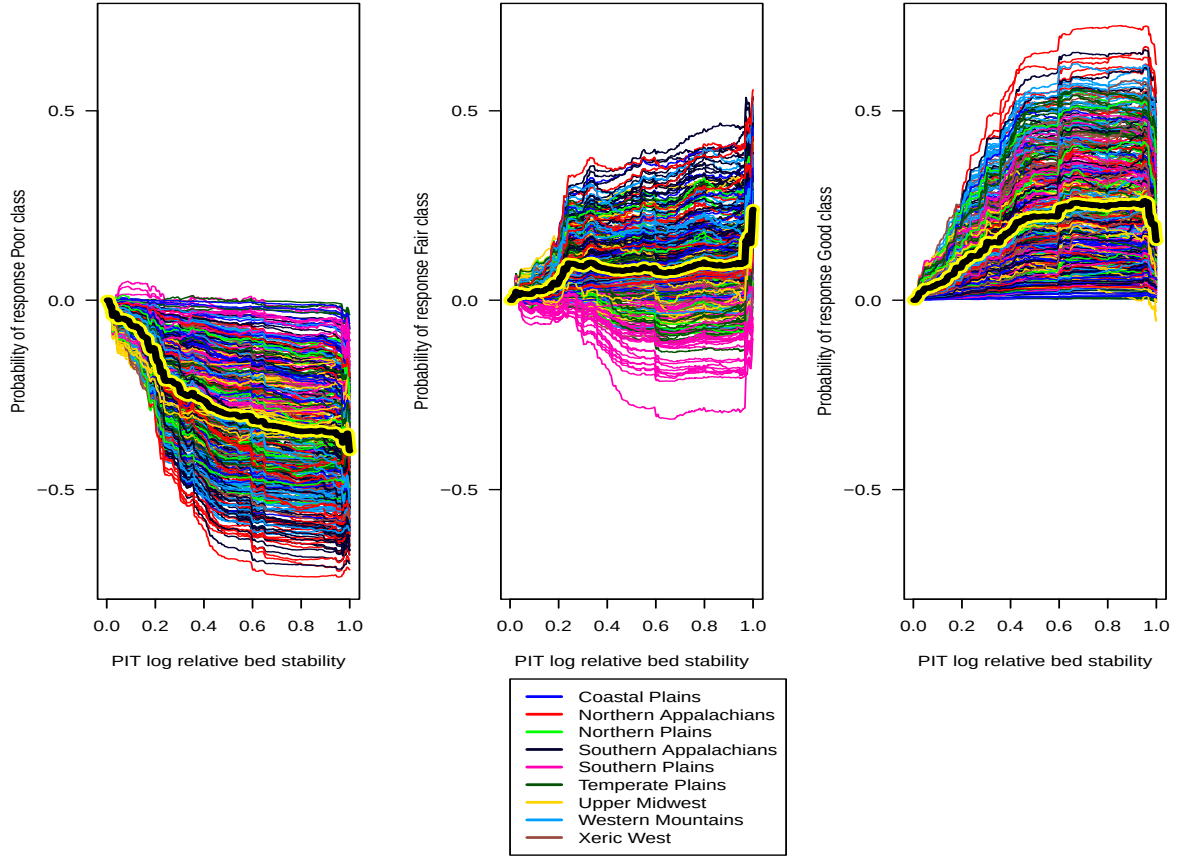


Figure S6: Centered marginal effect estimates of probability integral transformed (PIT) log relative bed stability on benthic MMI condition. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves). On the y-axis f_1 denotes the prediction of response class "Poor", f_2 represents prediction of response class "Fair" and f_3 equals prediction of response class "Good" as defined in Section 2.2. ICE curves were colored by ecoregion.

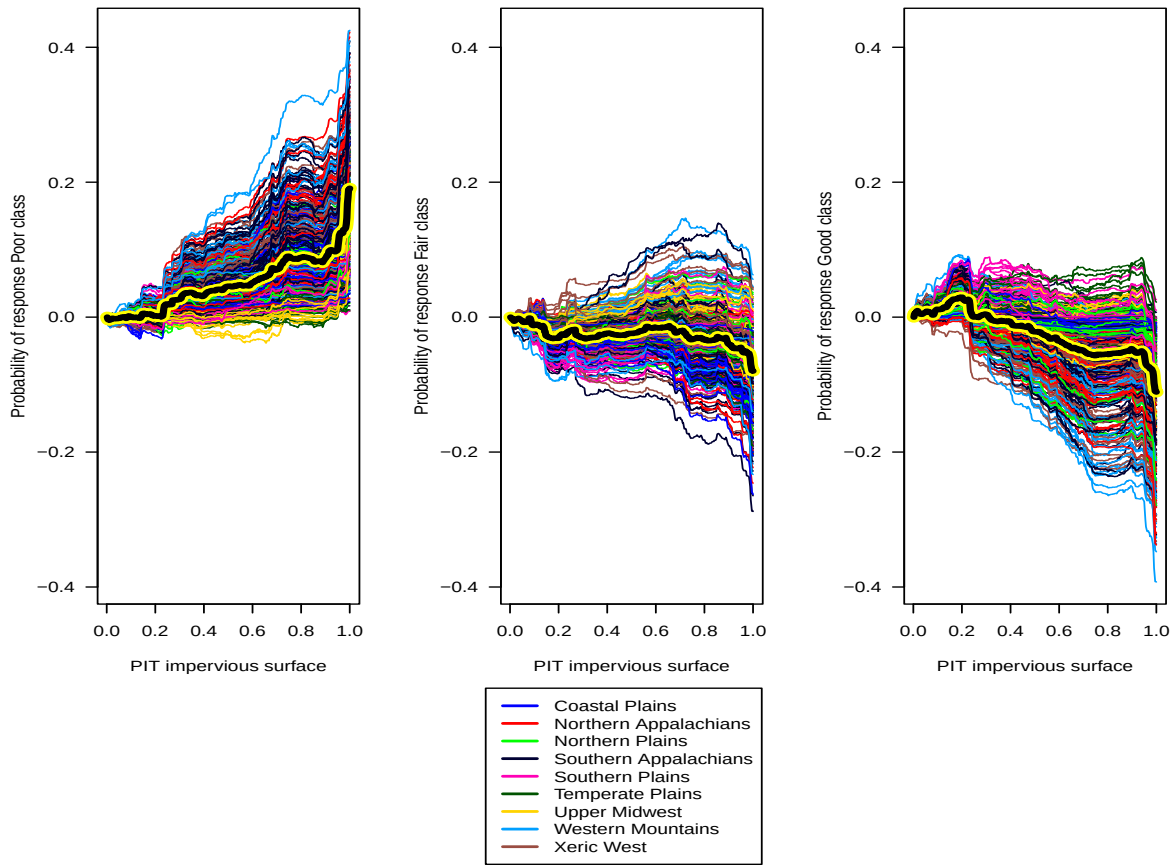


Figure S7: Centered marginal effect estimates of probability integral transformed (PIT) impervious surface on benthic MMI condition. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves). ICE curves were colored by ecoregion.

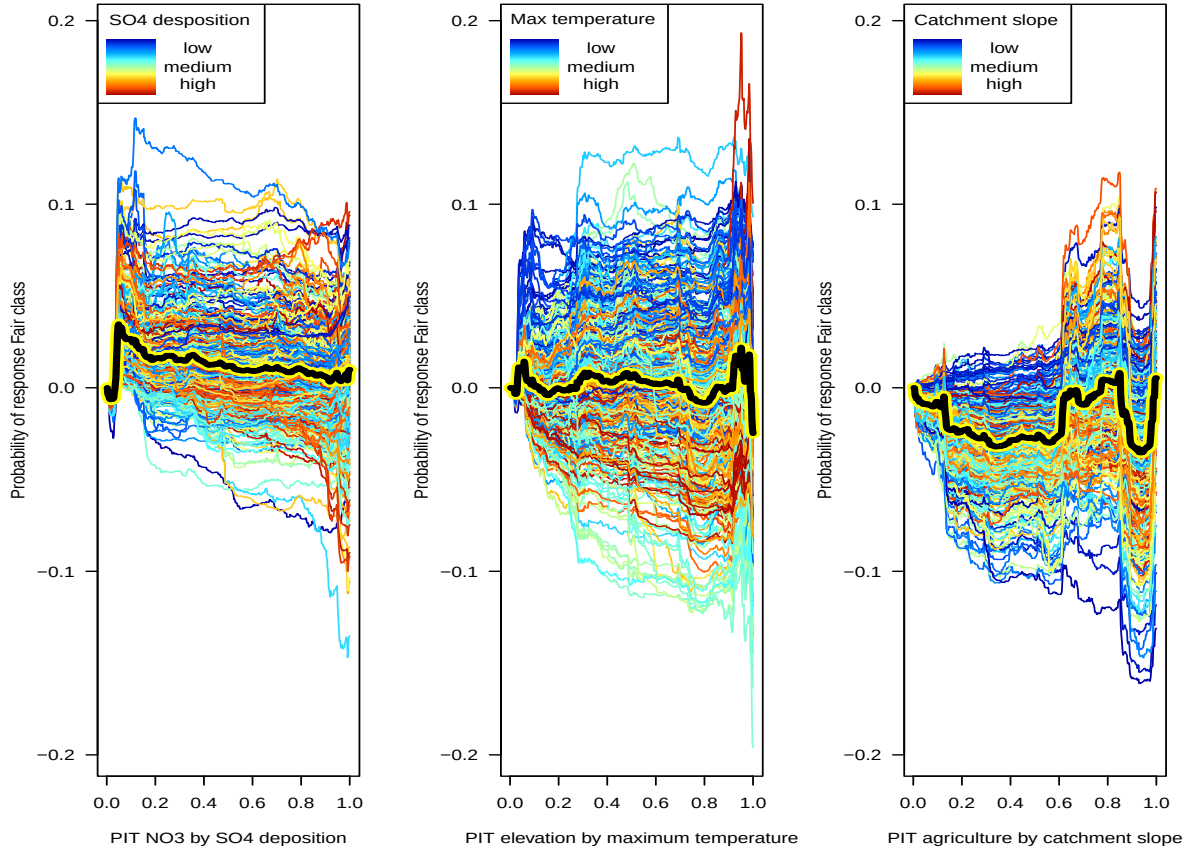


Figure S8: Centered ICE plots visualizing the three strongest significant two-way interaction effects as identified by parametric bootstrap test based on $\hat{H}_{j,k}^2$: NO_3 . Graphs were colored by probability integral transformed (PIT) SO_4 deposition (left panel), PIT elevation by maximal temperature (middle panel), PIT agriculture colored by catchment slope (right panel). Blue color corresponds to low and red color to high values. Here PDP highlights an interaction effect between elevation and maximum temperature regarding occurrence of Fair benthic MMI condition.

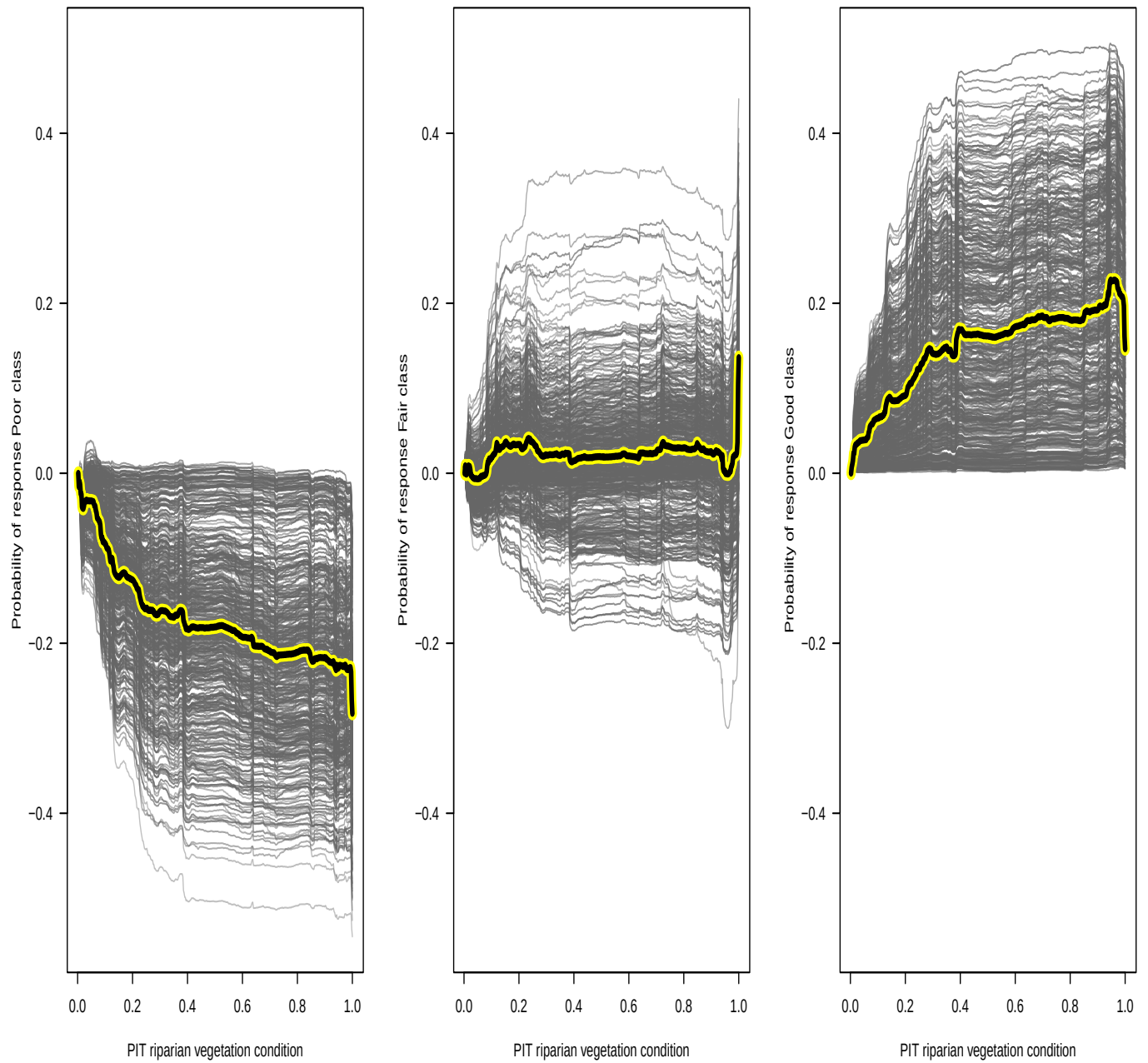


Figure S9: Marginal effect of probability integral transformed (PIT) riparian vegetation condition estimated benthic MMI condition distribution (centered). Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves).

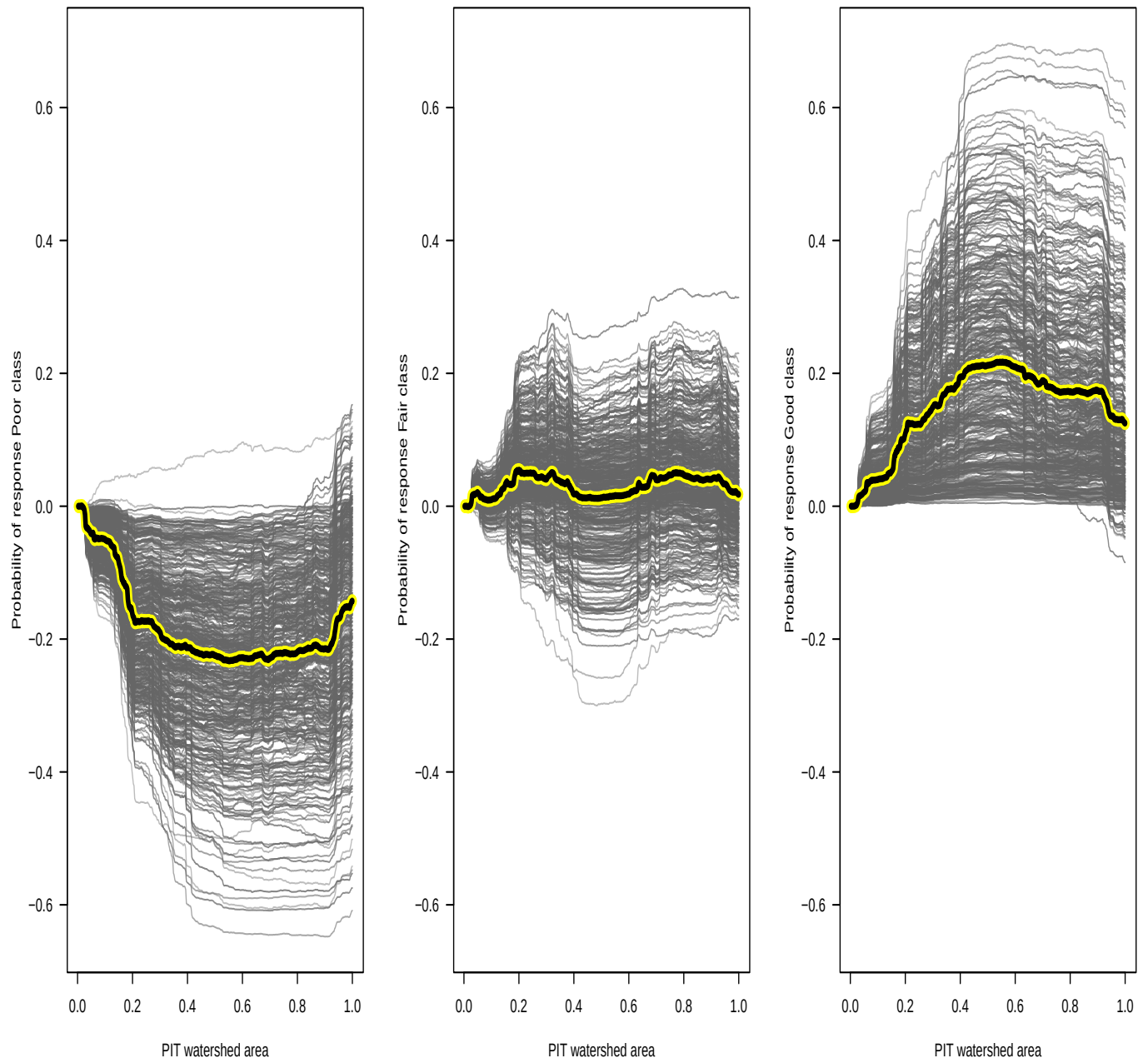


Figure S10: Centered marginal effect of probability integral transformed (PIT) watershed area on estimated benthic MMI condition distribution. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves).

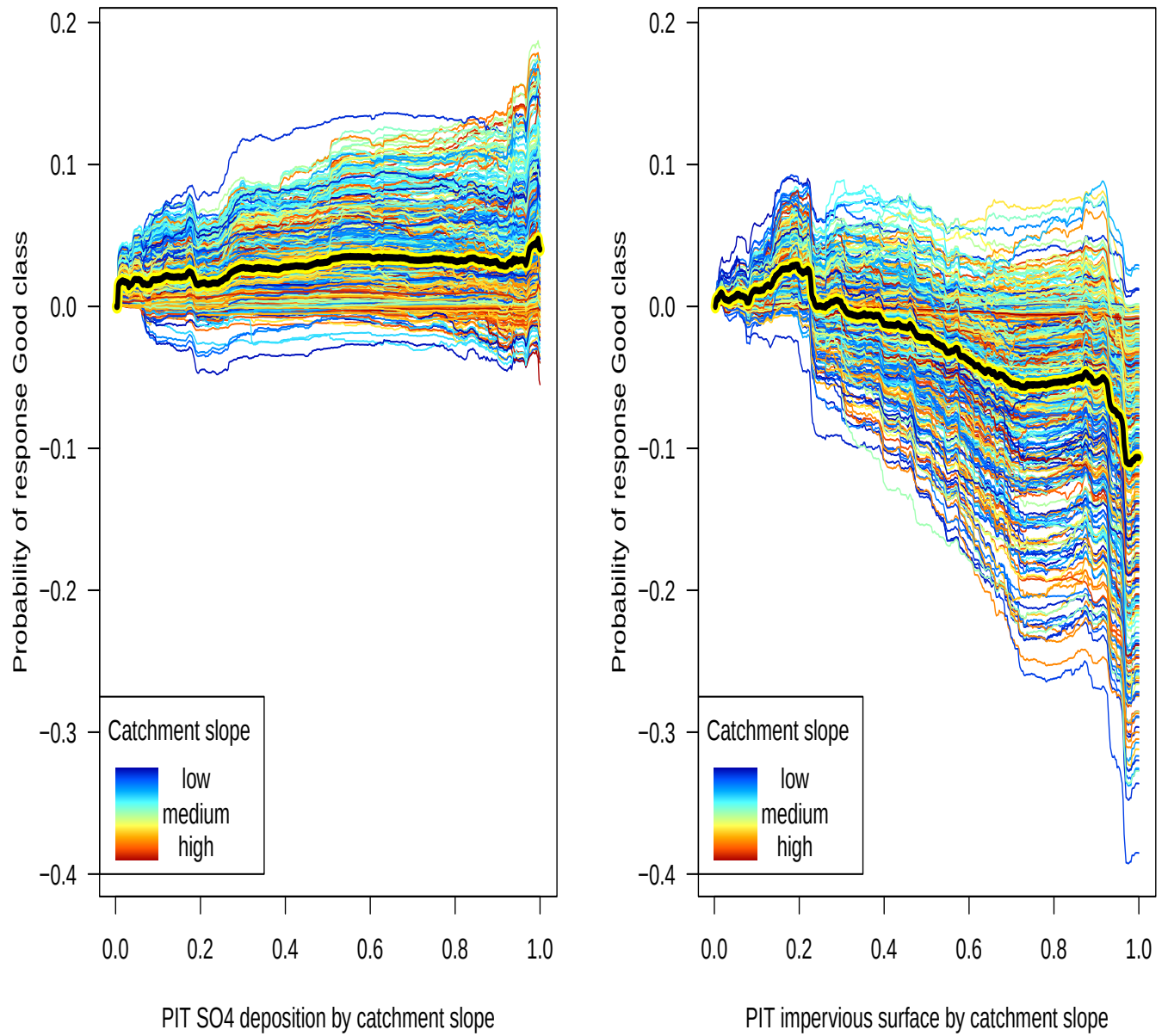


Figure S11: Centered marginal effect of probability integral transformed (PIT) impervious surface (right) on estimated benthic MMI condition distribution "Good". The colors represent low values (blue) to high values (red) of the second variable. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves).

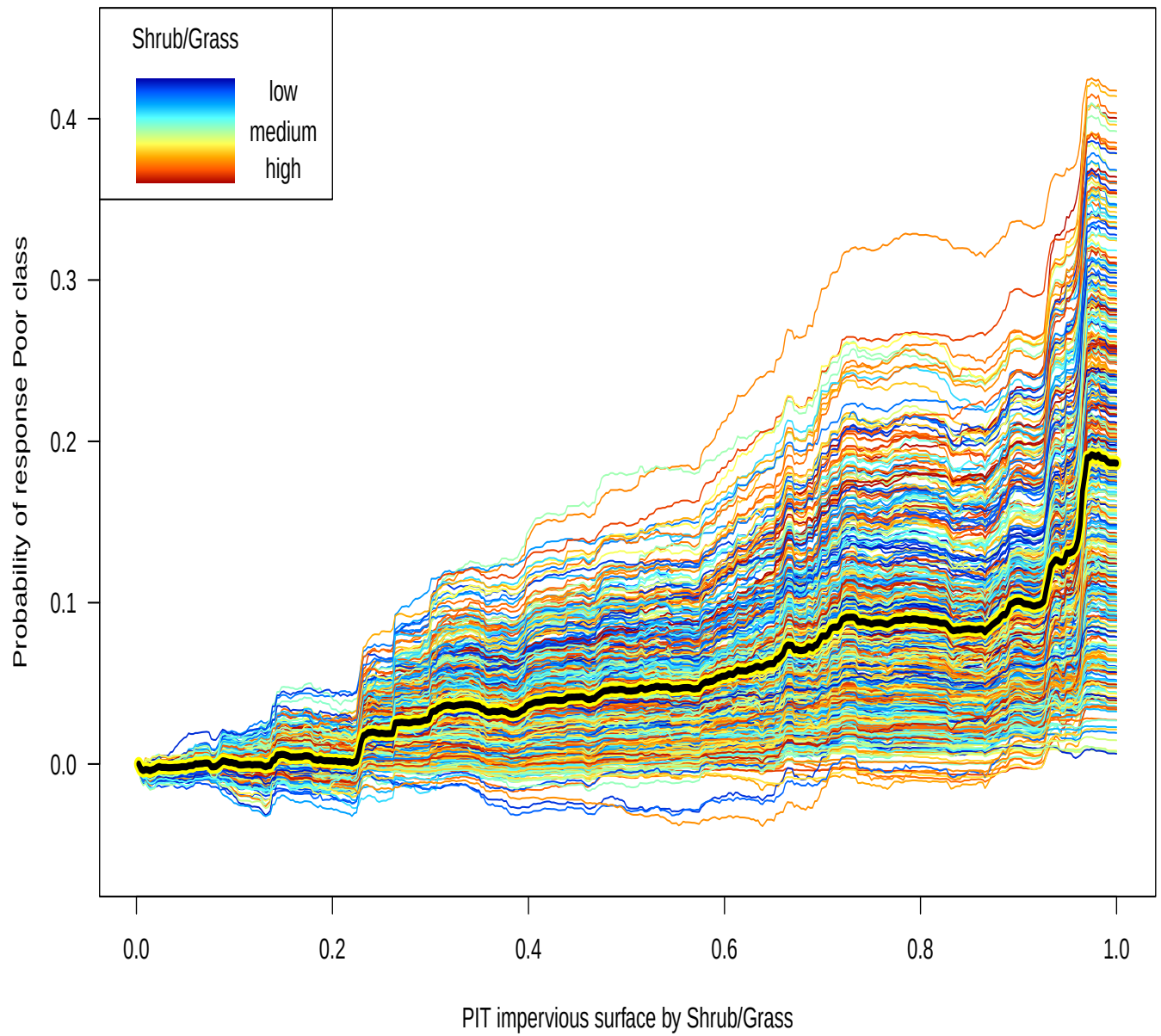


Figure S12: Centered marginal effect of probability integral transformed (PIT) impervious surface (right) on estimated benthic MMI condition distribution "Poor". The colors represent low values (blue) to high values (red) of the second variable. Each line represents one observation in the data set. Thick marked yellow lines correspond to the PDP curve (pointwise average over all curves).

1.4 Gaussian copula correlations

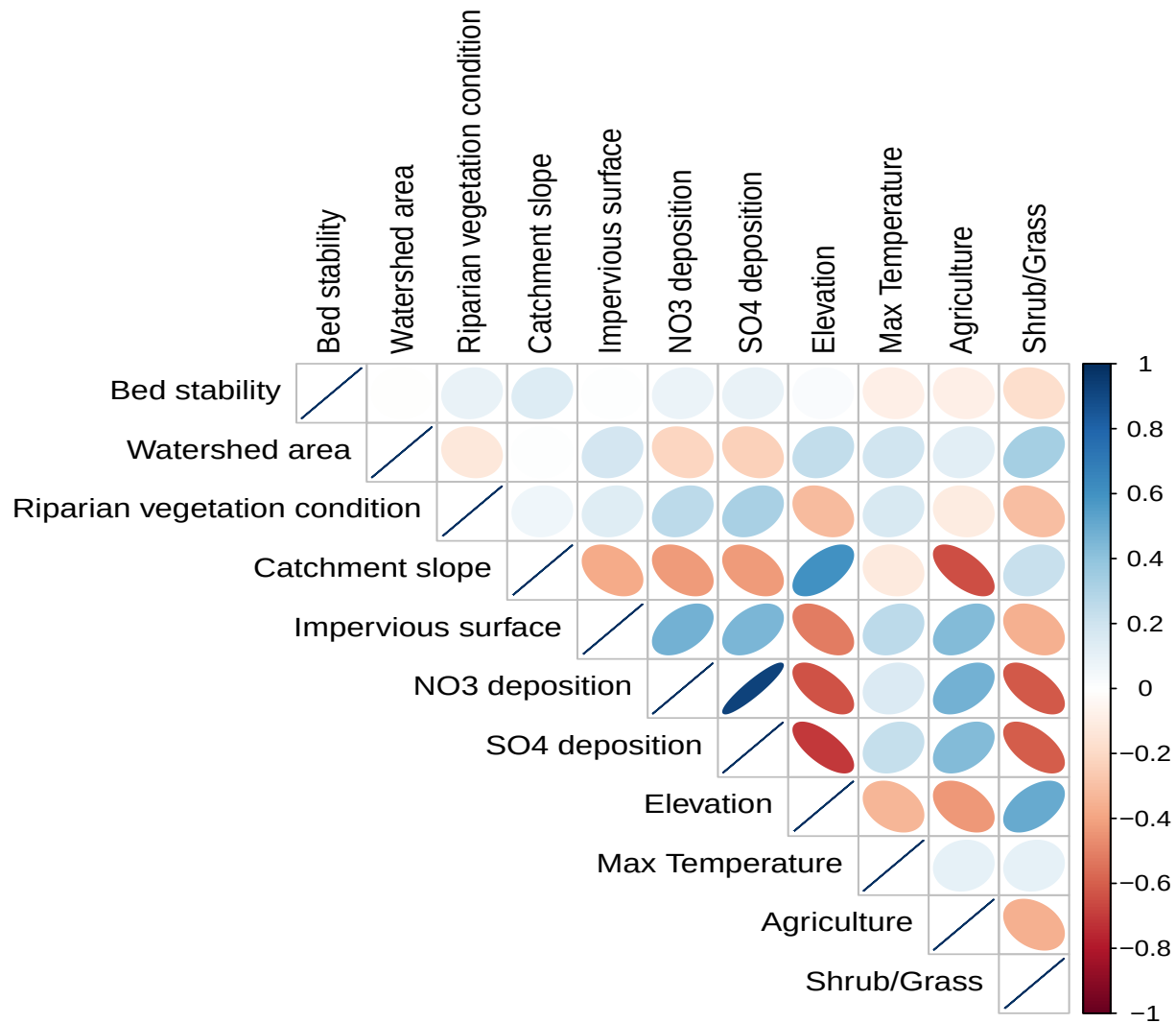


Figure S13: Correlation diagram of Gaussian copula. Estimated by applying empirical cumulative distribution transformation of the original data.

1.5 Final GBT model performance on complete data

Prediction / Data	Poor	Fair	Good
Poor	787	78	51
Fair	3	301	0
Good	21	37	516

Table S6: Confusion matrix of the final GBT model evaluated on the complete data set used in training the model. The accuracy was 89.41 %.

1.6 Parametric multinomial logit model

Variable	Response	OR	lowCI95	highCI95	adjustPValue
Ecoregion temperate plains	Good	4.4435	1.8488	10.6793	$2 * 10^{-2}$
Fish cover	Good	2.4141	1.4887	3.9148	$9 * 10^{-3}$
Riparian vegetation condition	Good	2.3375	1.3260	4.1208	$5 * 10^{-2}$
Log relative bed stability	Good	2.1097	1.6296	2.7312	$< 10^{-4}$
Fish cover	Fair	2.0537	1.2740	3.3106	$5 * 10^{-2}$
Log relative bed stability	Fair	2.0466	1.5949	2.6262	$< 10^{-4}$
SO4 deposition	Good	1.2697	1.1352	1.4200	$2 * 10^{-3}$
Catchment slope	Good	1.2008	1.1388	1.2661	$< 10^{-4}$
Catchment slope	Fair	1.0927	1.0345	1.1542	$3 * 10^{-2}$
Percent sandy soils	Good	1.0797	1.0384	1.1227	$4 * 10^{-3}$
Tree canopy	Good	1.0643	1.0367	1.0926	$< 10^{-4}$
Watershed area	Fair	1.0000	1.0000	1.0000	$9 * 10^{-3}$
Ecoregion western mountains	Good	0.0554	0.0167	0.1845	$< 10^{-4}$

Table S7: Summary of parametric multinomial logit model main effects (part I). Column *Variable* represents the names of the covariates. Column *Response* represents the probability of the benthic MMI condition class. Columns *lowCI95*, *highCI95* display the borders of the 95 % confidence intervals of the odds ratios in Column *OR*. Estimates were ordered by decreasing magnitude of the OR. The p-values were adjusted by controlling false discovery rate to 0.05 (Benjamini and Yekutieli, 2001) (column *adjustPValue*).

Variable1	Variable2	Response	OR	lowCI095	highCI095	adjustPValue
LRBS	Fish cover	Fair	1.6885	1.2739	2.2381	$8 * 10^{-3}$
LRBS	Fish cover	Good	1.6145	1.1895	2.1914	$4 * 10^{-2}$
Elevation	Agriculture	Good	1.0000	1.0000	1.0001	$3 * 10^{-2}$
Impervious surface	Watershed area	Fair	1.0000	1.0000	1.0000	$1 * 10^{-2}$
Ecoregion tpl	Watershed area	Fair	1.0000	1.0000	1.0000	$4 * 10^{-2}$
HDI	Watershed area	Good	1.0000	1.0000	1.0000	$5 * 10^{-3}$
Ecoregion sap	Watershed area	Good	0.9999	0.9998	1.0000	$4 * 10^{-2}$
Ecoregion nap	Watershed area	Fair	0.9999	0.9998	1.0000	$2 * 10^{-2}$
Ecoregion nap	Watershed area	Good	0.9999	0.9998	0.9999	$2 * 10^{-3}$
Max Temperature	Agriculture	Good	0.9971	0.9953	0.9990	$5 * 10^{-2}$
Percent sandy soils	Max Temperature	Good	0.9966	0.9944	0.9987	$3 * 10^{-2}$
SO4 deposition	Tree canopy	Good	0.9941	0.9919	0.9962	$< 10^{-4}$
HDI	Catchment slope	Good	0.9588	0.9325	0.9859	$5 * 10^{-2}$

Table S8: Summary of parametric multinomial logit model interaction effects (part II). Columns *Variable1*, *Variable2* represent the covariates within the second order interaction effect. Column *Response* represents the probability of the benthic MMI condition class. Columns *lowCI095*, *highCI095* display the borders of the 95 % confidence intervals of the odds ratios in Column *OR*. Estimates were ordered by decreasing magnitude of the OR. The p-values were adjusted by controlling false discovery rate to 0.05 (Benjamini and Yekutieli, 2001) (column *adjustPValue*). Abbreviations: Human disturbance index (HDI), log relative bed stability (LRBS), riparian vegetation condition (RVC), temperate plains (tpl), northern Appalachians (nap) and southern Appalachians (sap)

	Poor	Fair	Good
Poor	628	220	187
Fair	40	27	35
Good	143	169	345

Table S9: Confusion matrix of the parametric multinomial logit model evaluated by 10 fold outer stratified cross-validation. The accuracy was 55.47 %.

	Poor	Fair	Good
Poor	654	211	163
Fair	40	36	31
Good	117	169	373

Table S10: Confusion matrix of the parametric multinomial logit model evaluated on the complete data set used in training. The accuracy was 59.25 %.

2 Simulation

In this Section we conducted small simulation experiments to highlight how the presented methods of Section 2 of the manuscript perform in linear regression settings.

2.1 Comparison of PDP and ICE plots

Figure S14 illustrates how PDP and ICE curves perform while estimating the effects of an estimated GBT model (1000 trees, shrinkage 0.01, interaction depth=1) based on simulated data (sample size $n = 1000$) with a linear model $E(y|\mathbf{x}) = x_1 - x_2$ and independent covariates $x_1 \perp x_2$. The PDP curves correspond approximately to the linear effects of x_1, x_2 and the non-influential variable x_3 was identified correctly. Note that tree based models are essentially stepwise function approximators and therefore those methods have difficulties to discover true linear relationships.

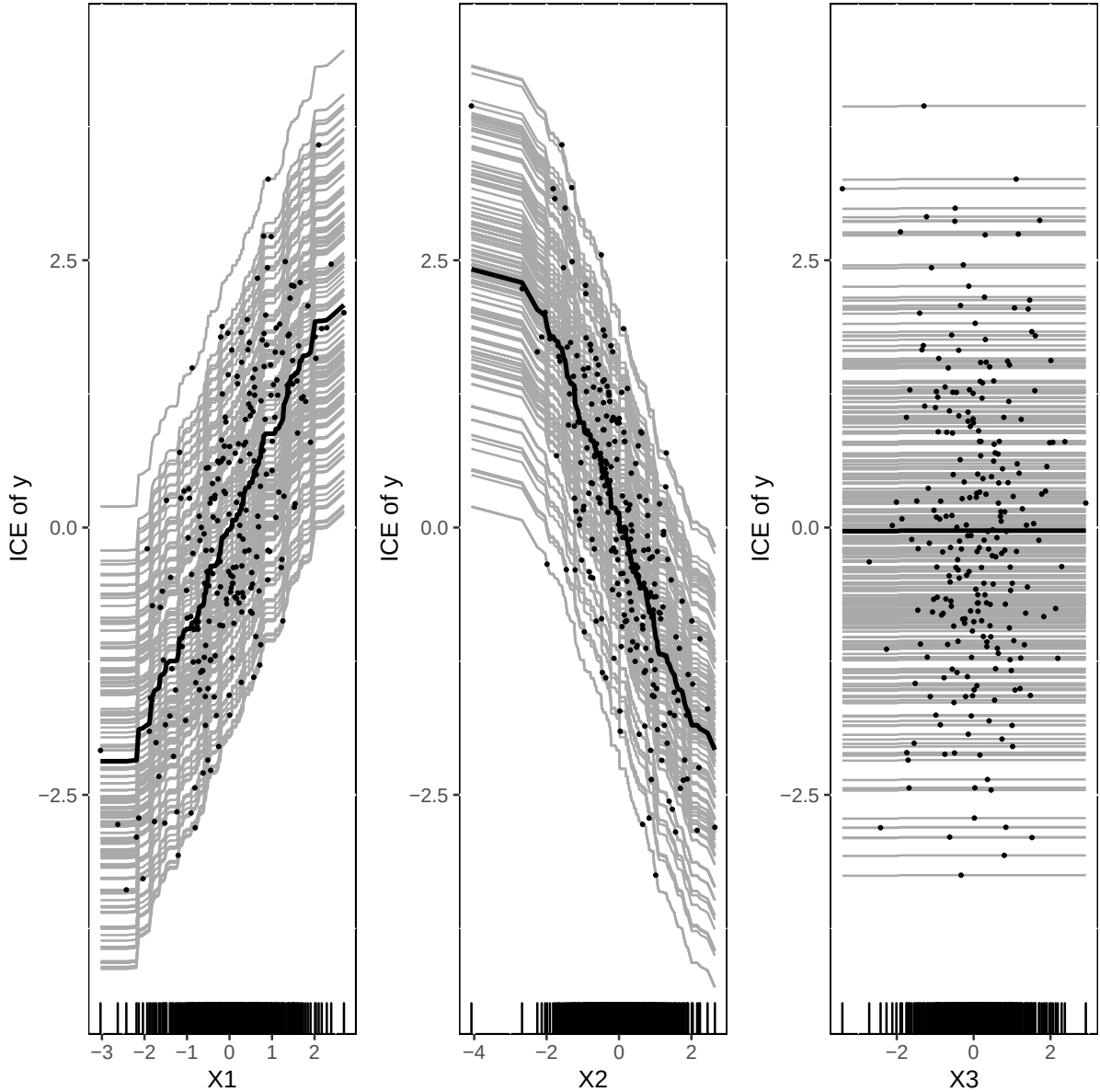


Figure S14: Simulation based on linear model $y = x_1 - x_2$ with no covariate correlations. The Figure displays ICE plots of the estimated boosted tree model (interaction depth of one and tuned number of iterations by 10-fold cross validation). The PDP curves approximately recover the linear effects and ICE curves are parallel due to only main effects.

2.2 Comparison of PDP and ALE plots

One of the major differences between PDP and ALE plots is that the former uses the complete covariate distribution without the covariate of interest and the latter conditions on the variable of interest. To illustrate how these different distributions differ consider a two-dimensional normal distribution with expectation zero and positive correlation of 0.8. We simulated $n = 1000$ samples in Figure S15.

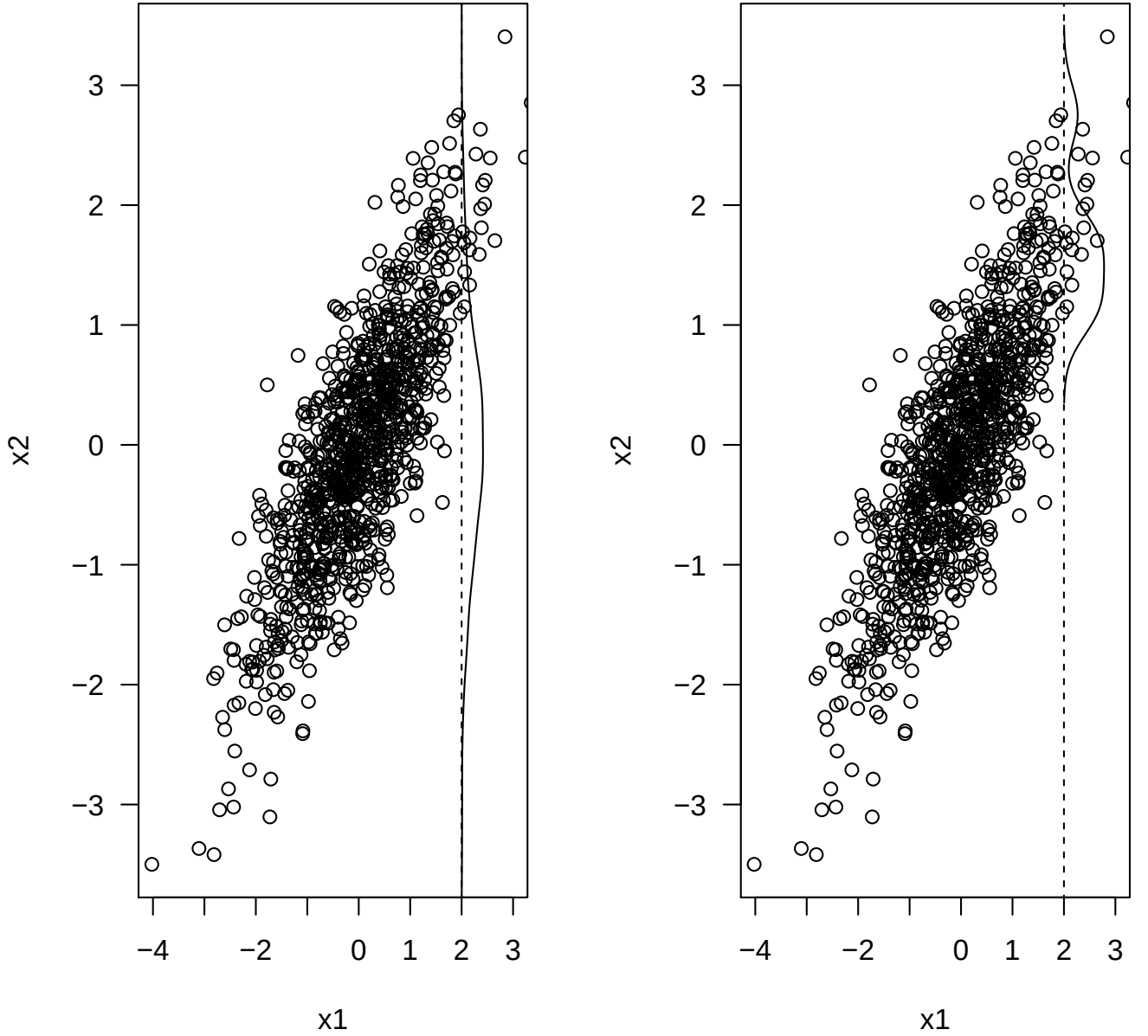


Figure S15: Simulated two-dimensional normal distribution with expectation zero and positive correlation of 0.8. The left part of the Figure shows the marginal distribution of $P(x_2)$, which is used to estimate PDP. For comparison the right side displays the conditional distribution of $P(x_2|x_1)$ that were used by ALE plot. In case of dependencies between covariates PDP averages over too low and high values of x_2 , which are not available in the data given $x_1 = 2$.

Clearly the conditional distribution $P(x_2|x_1)$ in the right part of the figure better captures the available data if $x_1 = 2$. The higher the correlations between x_2 and x_1 the more will the PDP and ALE plots differ.

To show an ALE example with an estimated GBT model we simulated a linear model based on the specification of the last Section 2.1 except for two differences: The variables are correlated $\rho(x_1, x_2) = 0.8$. Furthermore we modify the final prediction model: The predictions are equal to the estimated GBT model within the range of observed values of x_1, x_2 of the data set. However if observations lie within the set $S = \{(x_1 > 1.25 \cap x_2 < -1.25) \cup (x_1 < -1.25 \cap x_2 > 1.25)\}$ the prediction model changes the sign of the GBT predictions. The GBT model specification was the same as in the last paragraph. Due to the high correlation these cases are rare and no observations lie in this region with a sample size of $n = 1000$. ICE plots estimate the model behaviour outside the regions of available data (Figure S16). If the cases in set S never occur then ALE curves permit a more realistic interpretation. However if there is sample bias (e.g. underrepresented value combinations of covariates) and the goal is to detect those biases within the prediction model, then ICE plots are more suitable than ALE plots.

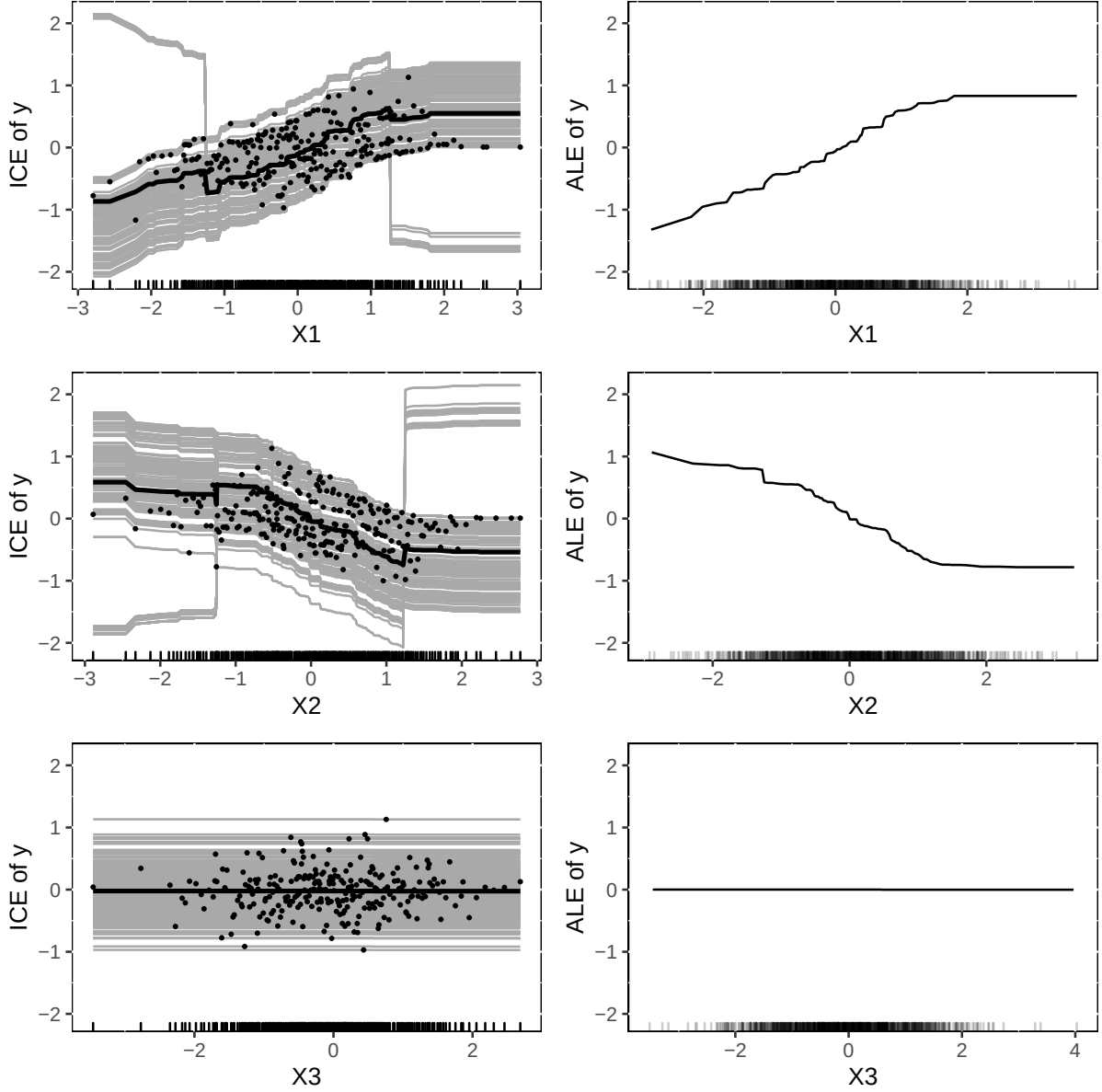


Figure S16: Simulation based on linear model $y = x_1 - x_2$ and correlations $\text{Corr}(x_1, x_2) = 0.75$. The right side of the Figure displays the estimated ICE plots of the GBT model. The prediction model changes the sign of the predictions within the set $S = \{(x_1 > 1.25 \cap x_2 < -1.25) \cup (x_1 < -1.25 \cap x_2 > 1.25)\}$. PDP curves include the jumps on the borders of set S in the estimated effects even if these cases are unrealistic (no observed cases). ALE plots do not contain these jumps and are more robust to changes in predictions outside the observed values of the data.

2.3 Marginal effects of covariates and influence of interaction effects

To compare how ICE and ALE plots change with increasing global interaction strength one data set was simulated based on a linear model with interaction effect $E(y|\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2$. Assumptions were the same as in the simulation example in Article Section 2.4.1. In this setting IAS and H^2 are equal, because there is only one

interaction effect. The different global interaction strength scenarios can be represented by the following coefficients:

$$\text{IAS} = 0 \Rightarrow \beta_1 = 1, \beta_2 = 1, \beta_{1,2} = 0 \quad (1)$$

$$\text{IAS} = 0.5 \Rightarrow \beta_1 = 1, \beta_2 = 1, \beta_{1,2} = \sqrt{2} \quad (2)$$

$$\text{IAS} = 1 \Rightarrow \beta_1 = 0, \beta_2 = 0, \beta_{1,2} = 1 \quad (3)$$

Next one data set was simulated (sample size $n = 1000$) for each global interaction strength and again a GBT model was estimated (1000 trees, shrinkage 0.01 and interaction depth of two). The first row in Figure S17 shows the estimated ICE and ALE plots in the absence of interaction effect and the linear effect approximately recovered in the GBT model. As expected the ICE plots show parallel lines in this case. Contrast this with average IAS in the second row. Higher values of both variables lead to an increase of variability of the non parallel ICE curves. The step downward slope in the plots with values $x_1 < 2$ seems to be an artifact because there are too few observations within this region. In the third row there are no main effects and the PDP and ALE curves correctly estimate that there is no positive or negative effect within $-2 < x_1, x_2 < 2$. Note that the ICE curves represent correctly that the variability of the interaction effect $\beta_{1,2}$ decreases if the values of x_1 or x_2 approach towards zero.

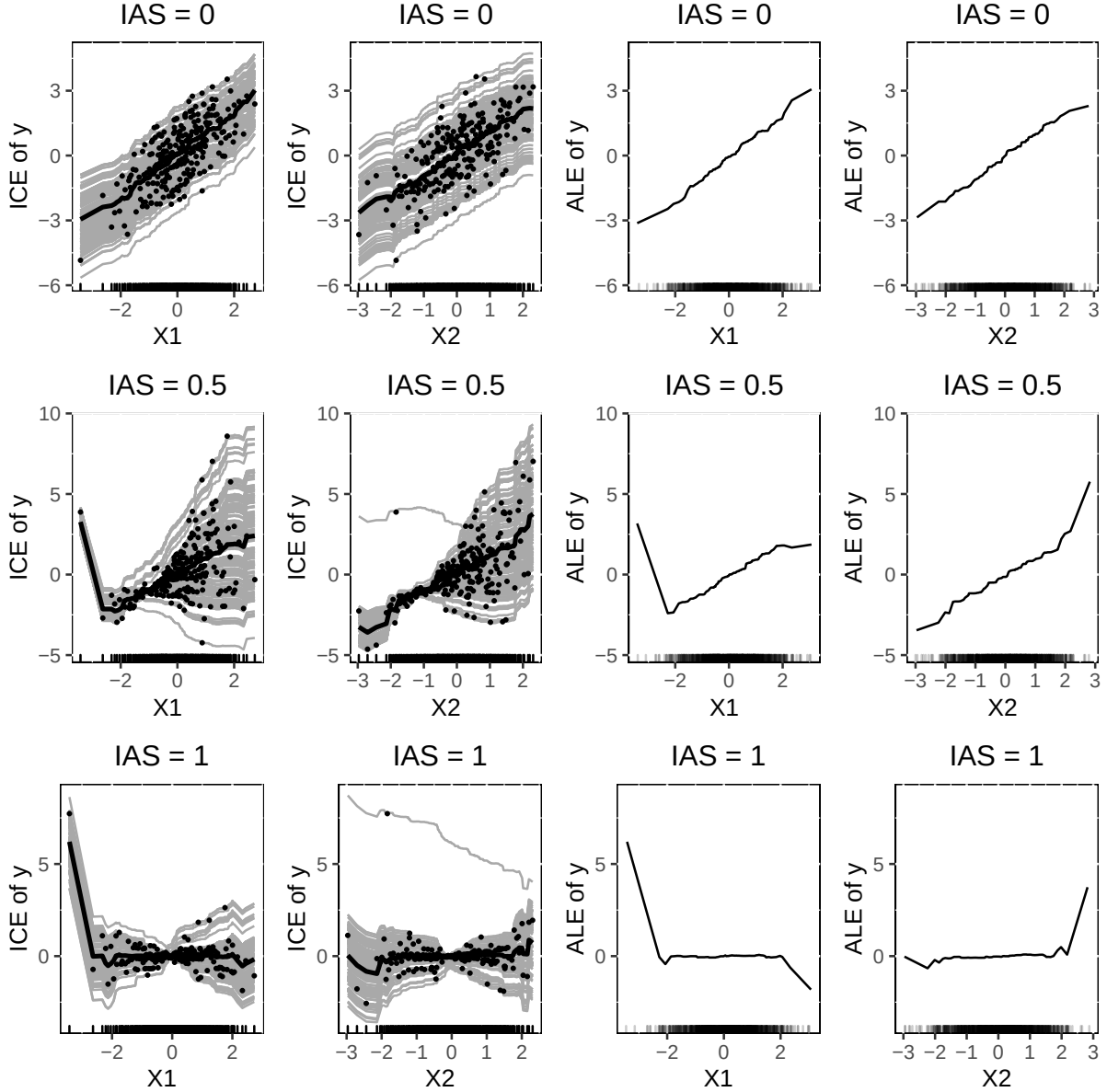


Figure S17: Comparison of ICE and ALE plots with varying interaction strength scenarios $IAS = 0, 0.5, 1$ in a linear regression setting. Both graphics estimate the effects of each covariate well within the dense region of available data. ICE plots show parallel lines in absence of interactions and ICE curves with varying slopes indicate presence of interaction effects.

Last but not least we wanted to give an example how the H^2 bootstrap test behaves in a linear regression context. One data set was simulated with a linear model with interaction effect $E(y|\mathbf{x}) = x_1 + x_2 + x_1x_2$ with the same assumptions as given in Section 2.4.1 of the article. The H^2 bootstrap test based on an estimated linear model $E(y|\mathbf{x}) = x_1 + x_2$ without interaction effects to simulate the null distribution and the correctly specified linear model (four independent standard normal variables, sample size $n = 100$, explained variance of 90 % and $m = 1000$ bootstrap repetitions). The evaluation of the interaction between x_1 and x_2 resulted in a significant result (p-value below 0.001) and the interaction effect between x_3 and x_4 was not significant (p-value=1). This small example demonstrates that the $H^2_{j,k}$ bootstrap test was able to detect the presence and non-presence of an interaction

effect within this scenario.