Analyses v2

Cape vs SWA

Ruan van Mazijk

2019-06-12

Preamble/outline

Here I layout the "new", second incarnation of the analyses as discussed over the course of May/June 2019, following the first draft of the manuscript.

To reiterate that manuscript, we hypothesise that the greater vascular plant species richness of the GCFR compared to that of the SWAFR is explained by the regions' difference in environmental heterogeneity.

The proposed "story" of questions for the analyses is as follows:

1. Is the GCFR more heterogeneous environmentally than the SWAFR, and does the scale of that heterogeneity differ to that of the SWAFR?
2. Do the regions differ w.r.t. the species richness of both HDS and QDS cells, and, for HDS cells' richness (S_{HDS}), does the explanatory power of mean QDS richness (S_{QDS}) and turnover (T_{QDS}) differ between the regions?
3. Does heterogeneity explain differences in richness and turnover between the regions?

1. Environmental heterogeneity & scale

Is the GCFR more heterogeneous environmentally than the SWAFR, and does the scale of that heterogeneity differ to that of the SWAFR?

In order to determine which region is more environmentally heterogeneous, and what scales heterogeneity is most pronounced, we calculated a measure of environmental heterogeneity at various spatial scales (namely: the base data resolution (0.05º x 0.05º), eighth- (EDS), quarter- (QDS), half- (HDS) and three-quarter-degree-squares (3QDS)).

Environmental "roughness" in both regions was calculated, in moving 3 x 3 cell windows, as the average absolute difference between cells and their (usually) 8 neighbours. Alternatively, for a focal cell x^*, the roughness is based on $x_1, x_2, \ldots, x_i, \ldots, x_8$ neighbour cells as:

$$Roughness(x^*) = f\begin{pmatrix} x_1 & x_2 & x_3 \\ x_4 & x^* & x_5 \\ x_6 & x_7 & x_8 \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n}|x^* - x_i|$$

In R, this is implemented this as follows:

```
roughness <- function(x) {
  raster::focal(x, matrix(1, nrow = 3, ncol = 3), function(x) {
    focal_cell <- x[5]
    neighbour_exists <- (!is.na(x)) & (!is.nan(x)) & (x != focal_cell)
    focal_exists <- (!is.na(focal_cell)) & (!is.nan(focal_cell))
    neighbour_cells <- x[neighbour_exists]
    if (focal_exists) {
```

```
      return(mean(abs(focal_cell - neighbour_cells)))
    } else {
      return(NA)
    }
  })
}
```

Following this, the various forms environmental heterogeneity were ordinated using principal component analysis (PCA), to summarise a major axis of heterogeneity in each region (Figure 1). Portions of the data matrices for each scale for these PCAs are shown in Table 1.

Both the actual environmental heterogeneity values and the principal component of heterogeneity were then compared between the GCFR and SWAFR using common language effect sizes ($CLES$). The $CLES$ of GCFR vs SWAFR heterogeneity values was regressed against the spatial scale at which it was calculated using simple linear regression (Figure 2, Table 2).

We can see that PDQ, NDVI, pH and, arguably, elevation are all consistently more heterogeneous in the GCFR than in the SWAFR, regardless of spatial scale (Figure 2). The GCFR is more heterogeneous at finer scales in terms of MAP, surface temperature, CEC and soil carbon (Figure 2). Notably, the GCFR is more pronouncedly heterogeneous at broad scales in terms of clay (Figure 2). In general (i.e. regarding PC1; Figure 2), the GCFR is more environmentally heterogeneous than the SWAFR, and particularly so at fine spatial scales.

Table 1: Portions of the data matrices used in the PCA for this section of the analysis, where roughness values were $log(x + 1)$-transformed to ensure normality.

region	Elevation	MAP	PDQ	Surface.T	NDVI	CEC	Clay	Soil.C	pH
GCFR	5.19	2.52	0.72	1.32	15.13	1.14	1.2	2.46	1.36
GCFR	5	2.7	0.61	1.16	15.01	1.11	1.11	1.74	1.83
GCFR	4.86	2.55	0.72	1.17	15.08	1.18	1.4	1.79	1.65
...	...	...	...	...	...	...	...	...	...
SWAFR	3.27	2.77	1.1	0.71	14.91	0.31	1.19	1.59	0.48
SWAFR	2.36	2.41	1.15	0.7	14.28	0.67	1.29	2.03	1.3
SWAFR	2.86	1.98	1.17	1.09	13.58	0.73	2.27	2.4	2.58

Table 2: Slopes, significances and R^2-values from regressions of $CLES$ against scale for each form of environmental roughness.

Variable	R^2	Slope	P_{slope}	
Elevation	0.891	0.044	0.016	*
MAP	0.874	-0.313	0.020	*
Surface.T	0.848	-0.330	0.026	*
Clay	0.902	0.243	0.013	*
Soil.C	0.967	-0.298	0.003	*
PC1	0.922	-0.172	0.010	*
CEC	0.735	-0.126	0.063	.
PDQ	0.253	0.010	0.387	
NDVI	0.193	0.032	0.459	
pH	0.037	-0.010	0.756	

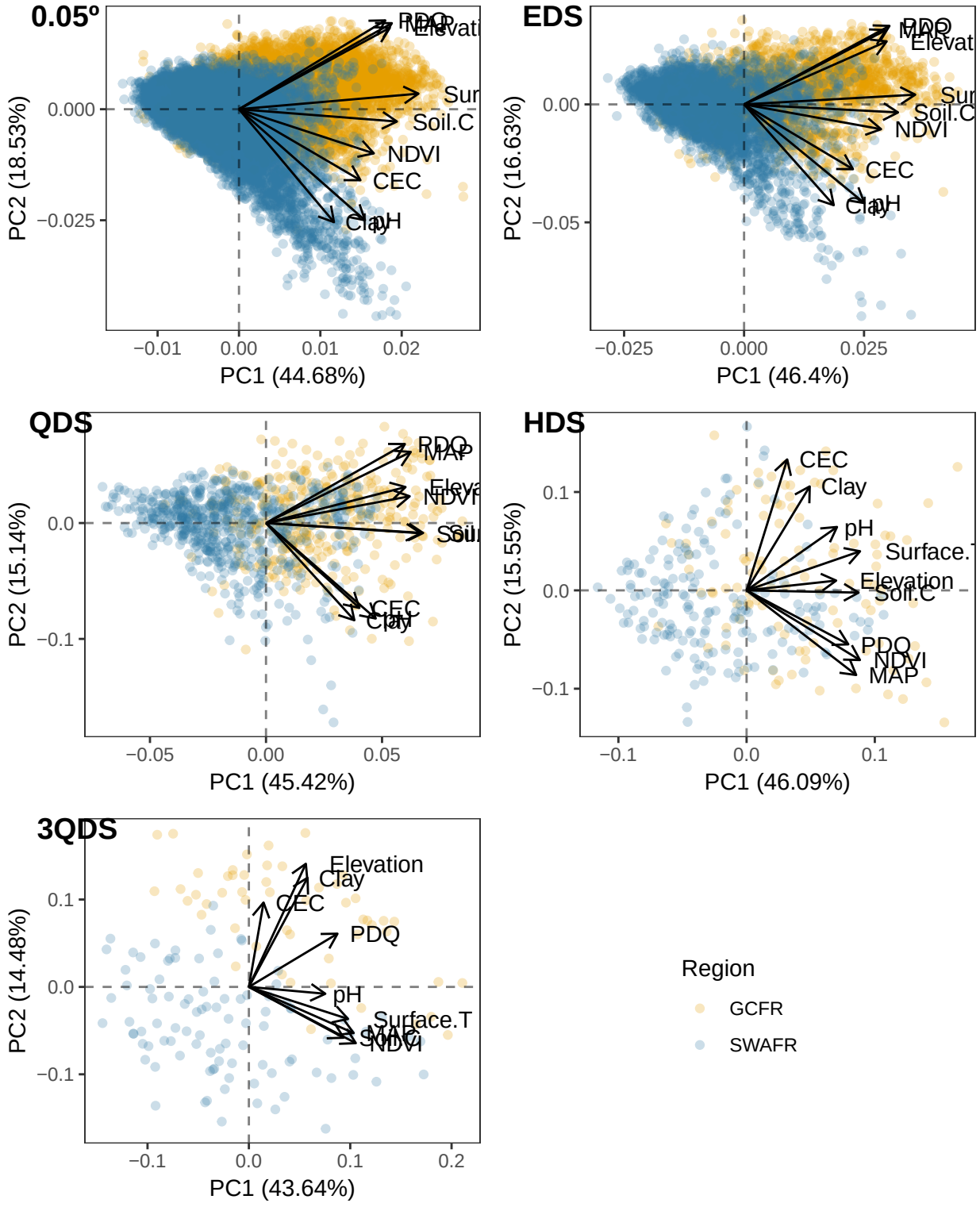


Figure 1: Scatter plots of the first and second principal components (PC1, PC2) of environmental heterogeneity following principal components analyses (PCAs) of the various forms of environmental heterogeneity, repeated at the five spatial scales. The proportion of variation accounted for by each axis is denoted in parentheses. Arrows (labelled) denote the rotational loading of a given form of environmental heterogeneity. Note, the signs of loadings on PC1 have been forced to be positive, while the signs of loadings on PC2 are arbitrary.

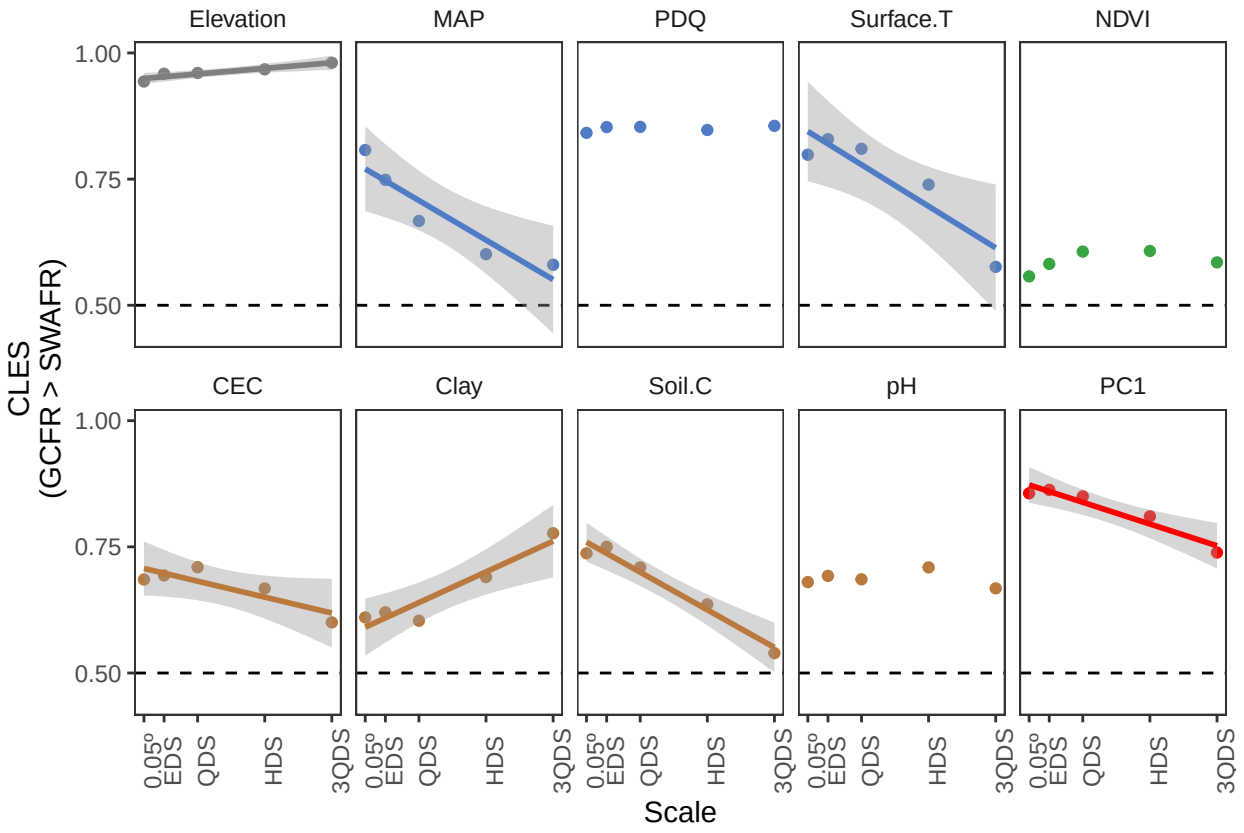


Figure 2: Regressions of the common language effect size ($CLES$) of various forms of environmental heterogeneity, and the first principal component thereof (PC1, see Figure 1). Only significant or marginally significant fits are plotted (Table 2).

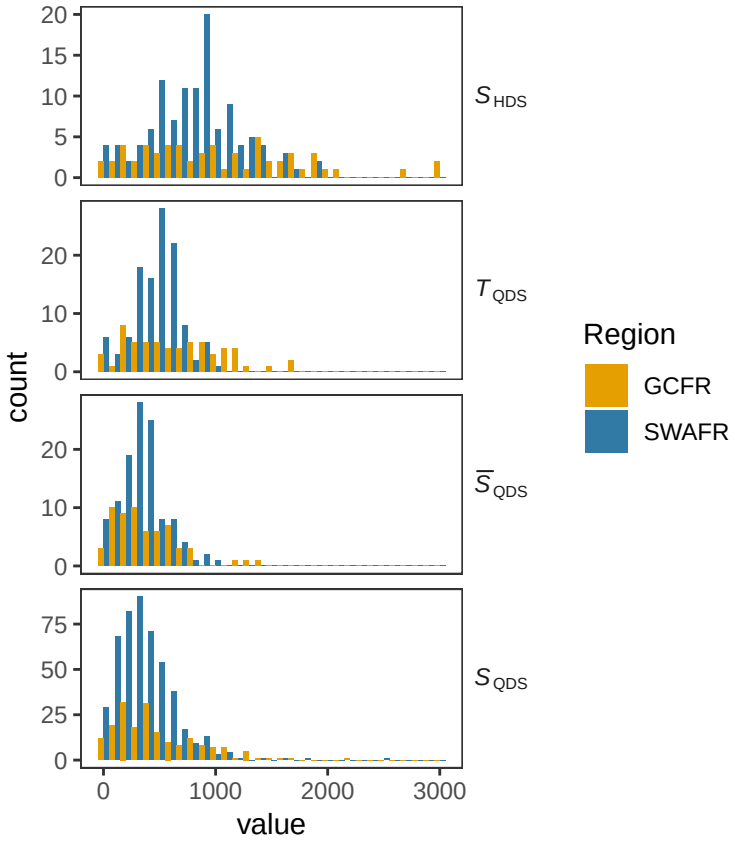


Figure 3: Distrbutions of the various species richness and turnover metrics in the GCFR and SWAFR.

2. Species richness & turnover

Do the regions differ w.r.t. the species richness of both HDS and QDS cells, and, for HDS cells' richness (S_{HDS}), does the explanatory power of mean QDS richness (S_{QDS}) and turnover (T_{QDS}) differ between the regions?

To tackle this question, I compare measures of species richness and turnover between the regions. Species richness at the HDS-scale (S_{HDS}) can be partitioned into the average richness of the constituent QDS in HDS ($\overline{S}_{\text{QDS}}$) and species turnover ($T_{\text{QDS}}$) defined[1] as:

$$T_{\text{QDS}} = S_{\text{HDS}} - \overline{S}_{\text{QDS}}$$

The distributions of these data are presented in Figure 3. To test for significant differences between GCFR and SWAFR values, I use Mann-Whitney U-tests and $CLES$ (Table 3), as most of the variables deviate significantly from normality (Table 4).

Additionally, a visualisation of how S_{HDS} is partitioned into $\overline{S}_{\text{QDS}}$ and T_{QDS} is presented in Figure 4.

We can conclude that broad scale species richness (i.e. that at the HDS scale) is more strongly driven by turnover between areas (i.e. QDS) than so in the SWAFR.

[1] following Whittaker's original additive definition: $\gamma = \alpha + \beta$

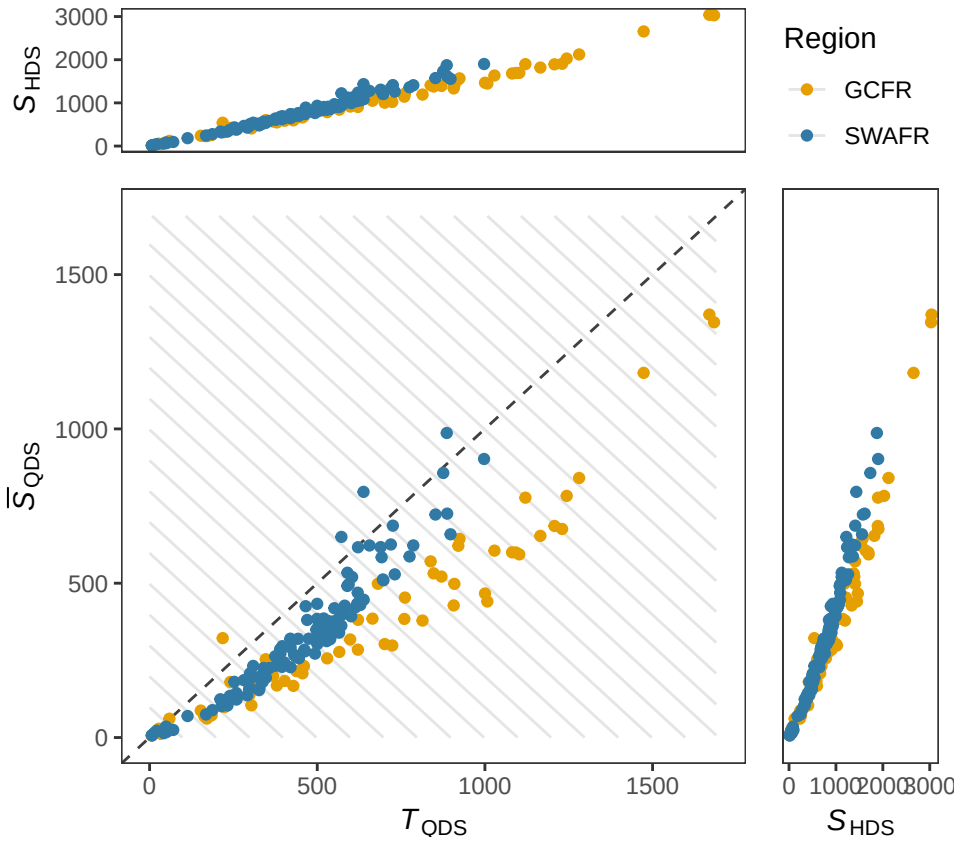


Figure 4: Scatter plot of $\overline{S}_{\mathrm{QDS}}$ and T_{QDS} with contour lines denoting the S_{HDS} that would arise as their sum (i.e. increasing from lower-left to upper-right). Additionally, the relationships between $\overline{S}_{\mathrm{QDS}}$ and S_{HDS} (upper panel) and T_{QDS} and S_{HDS} (rightmost panel) are illustrated in the panels flanking the central panel.

Table 3: Results of Mann-Whitney U-tests and the $CLES$ of GCFR vs SWAFR for various species richness and turnover metrics.

	$CLES$	P_U	
S_{HDS}	0.584	0.066	.
T_{QDS}	0.625	0.007	*
$\overline{S}_{\mathrm{QDS}}$	0.521	0.644	
S_{QDS}	0.576	0.002	*

Table 4: Results of Shapiro-Wilk tests of normality for various species richness and turnover metrics for the GCFR and SWAFR. [Significant deviations from normality: $P < 0.05$.]

	P_{GCFR}	P_{SWAFR}
S_{HDS}	0.008	0.168
T_{QDS}	0.075	0.100
$\overline{S}_{\mathrm{QDS}}$	≤ 0.001	0.002
S_{QDS}	≤ 0.001	≤ 0.001

3. Relating heterogeneity to species richness & turnover

Does heterogeneity explain differences in richness and turnover between the regions?

Here I fit various linear regressions of richness and turnover as functions of environmental heterogeneity.

[Interpretations to follow after meeting.]

3.1. Separate-regions models with combinations of variables

Table 5: Results of bi-directional stepwise multiple linear regressions of three richness and turnover responses in the against additive combinations of environmental heterogeneity variables. The stepwise regression procedure started with all variables included. (See Figure 5 for a graphical representation.)

Response	Predictor	Slope	P_{slope}	
GCFR S_{HDS}	Clay	368.640	0.005	*
GCFR S_{HDS}	MAP	619.047	0.000	*
GCFR S_{HDS}	pH	-229.074	0.108	
GCFR T_{QDS}	Clay	159.571	0.045	*
GCFR T_{QDS}	Elevation	79.913	0.163	
GCFR T_{QDS}	MAP	361.222	0.000	*
GCFR T_{QDS}	pH	-185.536	0.061	.
GCFR $\overline{S}_{\text{QDS}}$	Clay	67.586	0.082	.
GCFR $\overline{S}_{\text{QDS}}$	MAP	125.803	0.000	*
GCFR $\overline{S}_{\text{QDS}}$	NDVI	141.770	0.001	*
GCFR $\overline{S}_{\text{QDS}}$	pH	-147.157	0.001	*
GCFR $\overline{S}_{\text{QDS}}$	Soil.C	98.164	0.026	*
SWAFR S_{HDS}	CEC	-171.307	0.000	*
SWAFR S_{HDS}	Clay	89.809	0.045	*
SWAFR S_{HDS}	Elevation	177.141	0.000	*
SWAFR S_{HDS}	MAP	122.495	0.000	*
SWAFR S_{HDS}	NDVI	73.132	0.143	
SWAFR S_{HDS}	PDQ	130.328	0.004	*
SWAFR S_{HDS}	Surface.T	67.279	0.129	
SWAFR T_{QDS}	CEC	-92.644	0.000	*
SWAFR T_{QDS}	Clay	43.531	0.070	.
SWAFR T_{QDS}	Elevation	82.099	0.002	*
SWAFR T_{QDS}	MAP	60.518	0.000	*
SWAFR T_{QDS}	NDVI	59.978	0.018	*
SWAFR T_{QDS}	PDQ	70.153	0.004	*
SWAFR $\overline{S}_{\text{QDS}}$	CEC	-40.322	0.013	*
SWAFR $\overline{S}_{\text{QDS}}$	Clay	33.274	0.039	*
SWAFR $\overline{S}_{\text{QDS}}$	Elevation	43.084	0.011	*
SWAFR $\overline{S}_{\text{QDS}}$	MAP	97.198	0.000	*
SWAFR $\overline{S}_{\text{QDS}}$	PDQ	96.316	0.000	*
SWAFR $\overline{S}_{\text{QDS}}$	Surface.T	41.231	0.007	*

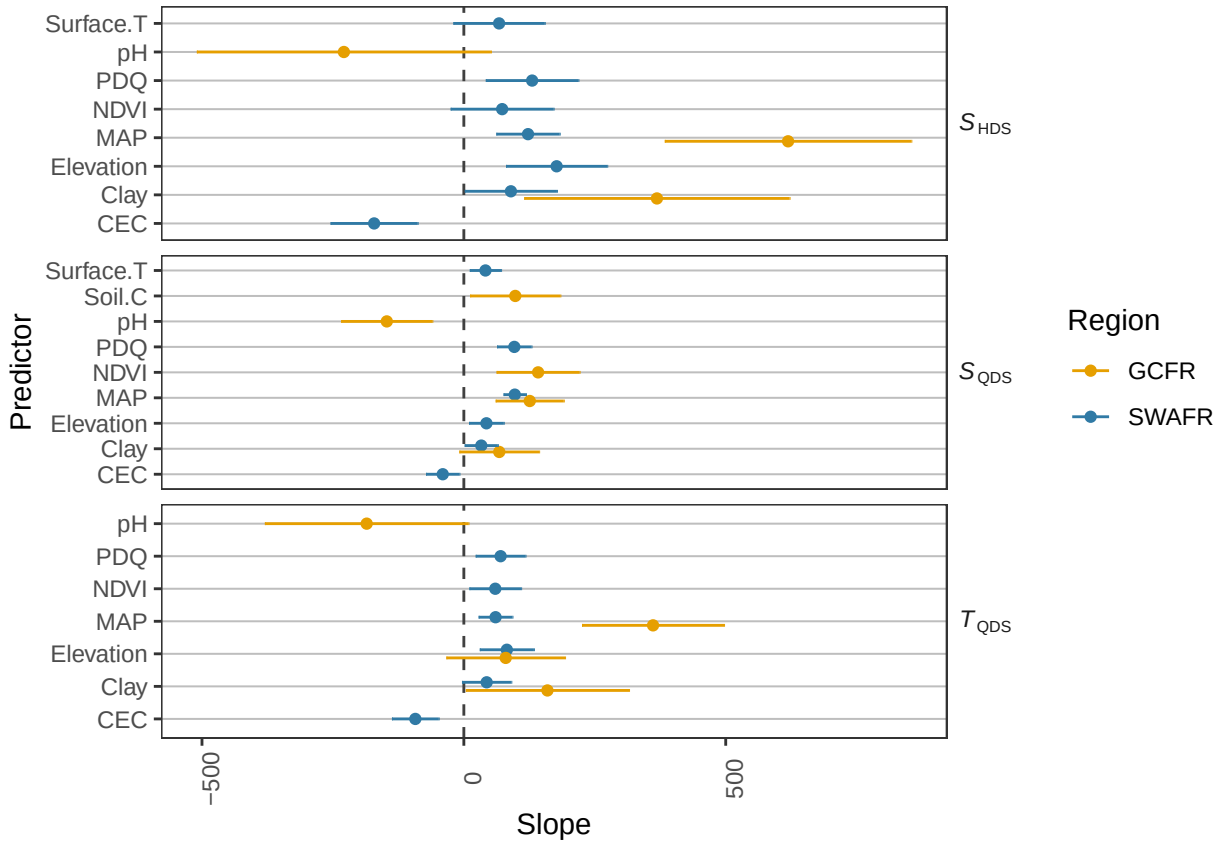


Figure 5: Slopes from Table 5, with error bars denoting 95% confidence intervals about each slope estimate.

Table 6: Adjusted R^2-values of the models in Table 5.

Response	$R^2{}_{adj.}$
GCFR S_{HDS}	0.347
GCFR T_{QDS}	0.372
GCFR $\overline{S}_{\text{QDS}}$	0.250
SWAFR S_{HDS}	0.516
SWAFR T_{QDS}	0.473
SWAFR $\overline{S}_{\text{QDS}}$	0.314

3.2. Combined-regions models with individual variables

3.2.1. Environmental heterogeneity variables

[Plots of the models referred to in Tables 7–9 can be discussed in a meeting.]

Table 7: Results of separate simple linear regressions of S_{HDS} against environmental heterogeneity variables with no region-term.

Predictor	R^2	P_{slope}	
CEC_roughness	0.013	0.134	
Clay_roughness	0.065	0.001	*
Elevation_roughness	0.161	0.000	*
MAP_roughness	0.294	0.000	*
NDVI_roughness	0.160	0.000	*
PDQ_roughness	0.143	0.000	*
pH_roughness	0.051	0.003	*
Soil.C_roughness	0.144	0.000	*
Surface.T_roughness	0.132	0.000	*

Table 8: Results of separate simple linear regressions of S_{HDS} against environmental heterogeneity variables with an additive region-term.

Predictor	R^2	P_{slope}		P_{region}	
CEC_roughness	0.050	0.630		0.011	*
Clay_roughness	0.101	0.002	*	0.010	*
Elevation_roughness	0.165	0.000	*	0.387	
MAP_roughness	0.294	0.000	*	0.744	
NDVI_roughness	0.170	0.000	*	0.140	
PDQ_roughness	0.143	0.000	*	0.957	
pH_roughness	0.075	0.027	*	0.035	*
Soil.C_roughness	0.151	0.000	*	0.236	
Surface.T_roughness	0.133	0.000	*	0.578	

Table 9: Results of separate simple linear regressions of S_{HDS} against environmental heterogeneity variables with an interaction-region-term.

Predictor	R^2	P_{slope}		P_{region}		$P_{slope:region}$	
CEC_roughness	0.081	0.024	*	0.588		0.017	*
Clay_roughness	0.102	0.141		0.017	*	0.619	
Elevation_roughness	0.173	0.002	*	0.313		0.203	
MAP_roughness	0.325	0.000	*	0.006	*	0.006	*
NDVI_roughness	0.173	0.001	*	0.502		0.460	
PDQ_roughness	0.144	0.003	*	0.693		0.565	
pH_roughness	0.091	0.009	*	0.880		0.086	
Soil.C_roughness	0.151	0.003	*	0.499		0.817	
Surface.T_roughness	0.133	0.006	*	0.594		0.947	

3.2.2. PC1 models

Here, I present my findings with raw R-code, because I don't have the time to format it neatly.

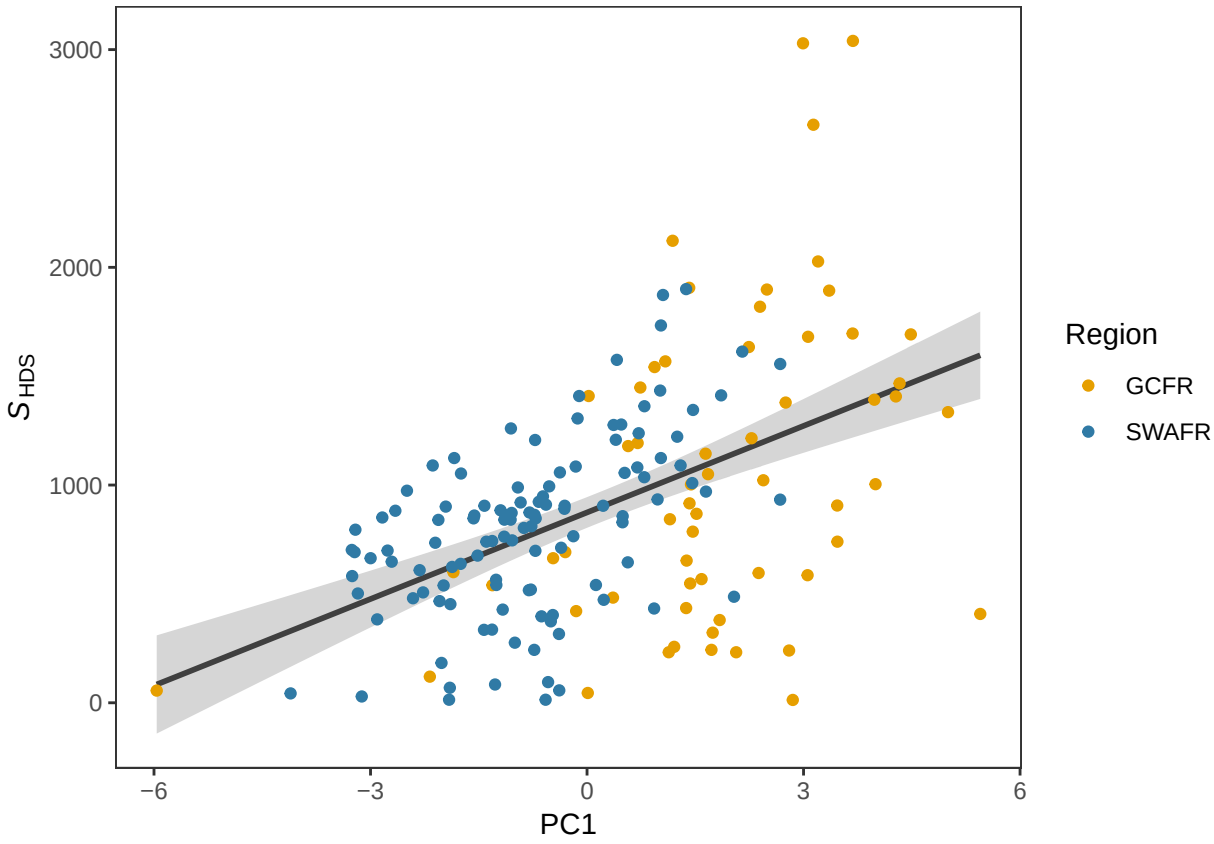
```
# Richness (HDS)

m1 <- lm(HDS_richness ~ PC1,          HDS)
m2 <- lm(HDS_richness ~ PC1 + region, HDS)
m3 <- lm(HDS_richness ~ PC1 : region, HDS)
m4 <- lm(HDS_richness ~ PC1 * region, HDS)
AIC(m1, m2, m3, m4)
```

```
##    df      AIC
## m1  3 2655.279
## m2  4 2655.022
## m3  4 2657.217
## m4  5 2656.999
```

```
summary(m1)
```

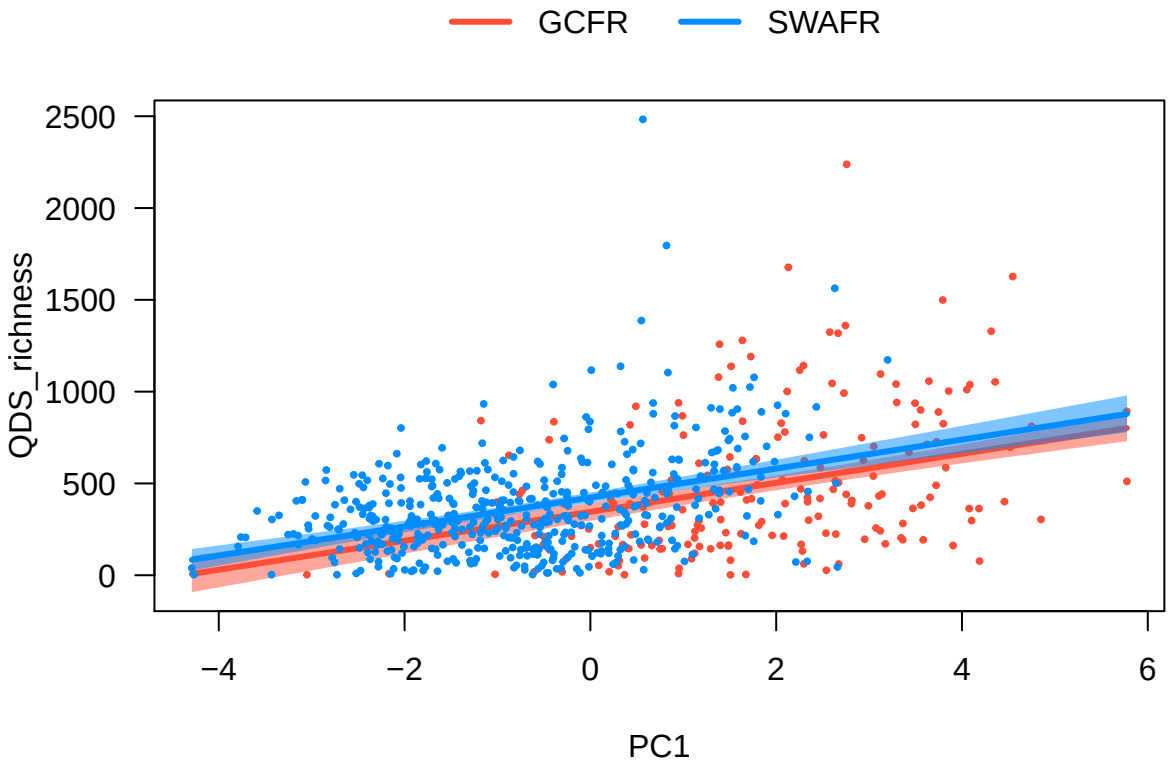
```
##
## Call:
## lm(formula = HDS_richness ~ PC1, data = HDS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1238.96  -293.58    40.94   243.17  1758.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   874.17      35.71  24.478  < 2e-16 ***
## PC1           132.51      17.86   7.419 5.08e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 471.6 on 173 degrees of freedom
## Multiple R-squared:  0.2414, Adjusted R-squared:  0.237
## F-statistic: 55.05 on 1 and 173 DF,  p-value: 5.078e-12
```

```
ggplot(HDS, aes(PC1, HDS_richness)) +
  geom_smooth(method = lm, colour = "grey25") +
  geom_point(aes(colour = region)) +
  ylab(bquote(italic("S")["HDS"])) +
  scale_colour_manual(name = "Region", values = my_palette)
```



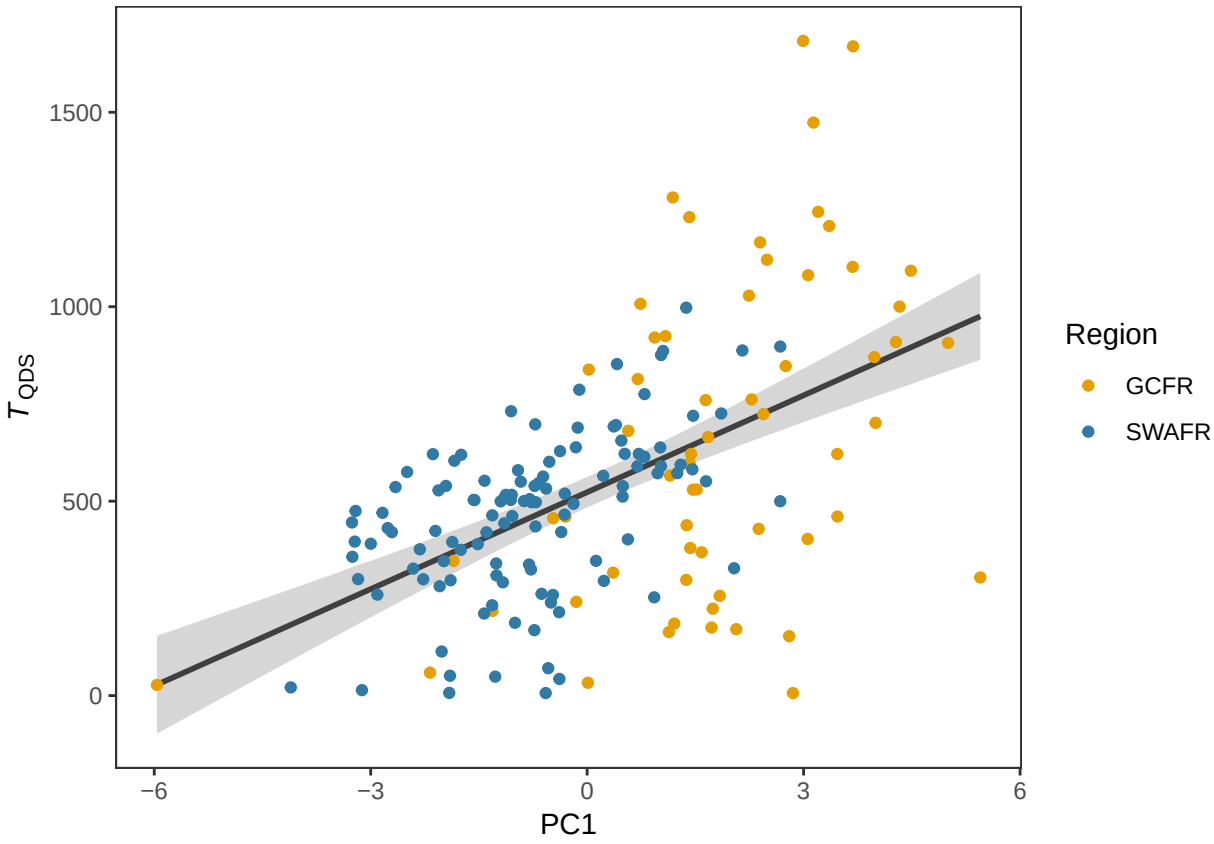

```
##    df      AIC
## m1  3 9523.008
## m2  4 9518.832
## m3  4 9524.919
## m4  5 9518.171
```

```
##
## Call:
## lm(formula = QDS_richness ~ PC1 + region, data = QDS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -599.12 -197.92  -40.61  138.70 2015.47
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  345.502     24.800  13.931   <2e-16 ***
## PC1           78.912      7.599  10.385   <2e-16 ***
## regionSWAFR   77.354     31.122   2.485   0.0132 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 284.1 on 670 degrees of freedom
## Multiple R-squared:  0.1666, Adjusted R-squared:  0.1641
## F-statistic: 66.96 on 2 and 670 DF,  p-value: < 2.2e-16
```




```
##    df      AIC
## m1  3 2450.583
## m2  4 2452.381
## m3  4 2451.624
## m4  5 2453.082
```

```
##
## Call:
## lm(formula = add_turnover ~ PC1, data = HDS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -753.14 -164.52   21.99  136.97  912.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  522.951     19.899  26.280  < 2e-16 ***
## PC1           83.016      9.952   8.342 2.22e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 262.8 on 173 degrees of freedom
## Multiple R-squared:  0.2869, Adjusted R-squared:  0.2827
## F-statistic: 69.59 on 1 and 173 DF,  p-value: 2.225e-14
```



3.3. Combined-regions models with combinations of variables

[To be discussed in meeting.]