Supporting Information

For “Heterogeneity and species richness”

Ruan van Mazijk, Michael D. Cramer and G. Anthony Verboom

# Species occurrence data cleaning

Firstly, we retained only records identified to the species level, and ignored intraspecific taxa. This resulted in the retention of 14,147 and 8,912 unique species names for the GCFR and SWAFR, respectively. The R package “taxize” (Chamberlain et al., 2016) was then used to query each species name against two major taxonomic databases, the Global Name Resolver (GNR) and the Taxonomic Name Resolution Service (TNRS; Boyle et al., 2013). Where either or both databases returned a match for a name, the name was retained; where not, it was excluded. Although the number of species thus excluded is high (GCFR: 692; SWAFR: 1,171), the geographically-random distribution of the records associated with these names suggests that exclusion of these names will not significantly influence spatial patterns of species richness.

In order to ensure that no species were listed under multiple synonyms, the retained names were then queried against the Tropicos and Integrated Taxonomic Information System (ITIS) for known synonyms, again using “taxize”. We then removed all records of species identified as non-native, using lists of invasive plants for South Africa and Australia from the IUCN’s Global Invasive Species Database (<http://www.iucngisd.org/gisd/>).

Finally, we removed species with fewer than five total collection records in total, in order to exclude collections with potentially low-confidence identifications. This, and the exclusion of occurrence data originating from coastal pixels at the 0.05° resolution, brought the total number of species in each region down to 9,419 and 6,696 in the GCFR and SWAFR respectively.

# Tables

**Table S1:** Georeferenced environmental data1 and vascular plant species occurrence data sources used in this study. Data were acquired for the GCFR and SWAFR, with the temporal extent of data products used described where applicable.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset(s) | Source | Temporal extent | Citation(s) |
| Plant species occurrences | GBIF |  | GBIF (2017a,b) |
| Elevation | SRTM (v2.0) |  | Farr et al. (2007) |
| NDVI, Surface T | MODIS (v006) | Feb. 2000 to Apr. 2017 | NASA (2017a,b) |
| MAP, PDQ | CHIRPS (v2.0) | Jan. 1981 to Feb. 2017 | Funk et al. (2015) |
| CEC, clay, soil C, pH | SoilGrids250m |  | Hengl et al. (2017) |

1 Abbreviations are as follows: NDVI, normalized difference vegetation index; T, temperature; MAP, mean annual precipitation; PDQ, precipitation in the driest quarter; CEC, cation exchange capacity; C, carbon.

**Table S2:** Analyses of variances (ANOVAs) of the three multiple regression models of vascular plant species richness, across the GCFR and SWAFR, including species richness hotspots (Figure 5). The variables in each model are arranged in descending order according to their proportion of variance explained. The significance1 (where applicable) of each variable’s contribution to each model is also shown. Abbreviations are as in Table S1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Response | Term | Variance explained | | |
| (a) *S*QDS | (Residuals) | 0.75 |  |  |
|  | MAP | 0.14 |  | \*\*\* |
|  | Elevation | 0.06 |  | \*\*\* |
|  | Region | 0.01 |  | \*\*\* |
|  | NDVI | 0.01 |  | \*\*\* |
|  | PDQ | 0.01 |  | \*\* |
|  | pH × Region | 0.01 |  | \* |
|  | PDQ × Region | 0.01 |  | \* |
|  | pH | <0.01 |  | ~ |
|  | NDVI × Region | <0.01 |  | ~ |
|  | MAP × Region | <0.01 |  | ~ |
|  | CEC × Region | <0.01 |  |  |
|  | CEC | <0.01 |  |  |
| (b) *S*HDS | (Residuals) | 0.62 |  |  |
|  | MAP | 0.19 |  | \*\*\* |
|  | NDVI | 0.05 |  | \*\*\* |
|  | Elevation | 0.04 |  | \*\*\* |
|  | Clay | 0.02 |  | \*\* |
|  | PDQ × Region | 0.01 |  | \* |
|  | PDQ | 0.01 |  | \* |
|  | pH × Region | 0.01 |  | ~ |
|  | Elevation × Region | 0.01 |  | ~ |
|  | CEC × Region | 0.01 |  | ~ |
|  | MAP × Region | 0.01 |  |  |
|  | pH | 0.01 |  |  |
|  | Soil C × Region | <0.01 |  |  |
|  | Region | <0.01 |  |  |
|  | CEC | <0.01 |  |  |
|  | Soil C | <0.01 |  |  |
| (c) *S*DS | (Residuals) | 0.31 |  |  |
|  | MAP | 0.27 |  | \*\*\* |
|  | Elevation | 0.12 |  | \*\*\* |
|  | Clay | 0.05 |  | \*\* |
|  | Clay × Region | 0.05 |  | \*\* |
|  | PDQ | 0.04 |  | \* |
|  | Soil C | 0.04 |  | \* |
|  | Surface T × Region | 0.04 |  | \* |
|  | Elevation × Region | 0.03 |  | \* |
|  | pH × Region | 0.03 |  | \* |
|  | pH | 0.02 |  |  |
|  | PDQ × Region | 0.01 |  |  |
|  | Surface T | <0.01 |  |  |
|  | Region | <0.01 |  |  |

1 Represented as follows: \*\*\*, *P* < 0.001; \*\*, *P* < 0.01; \*, *P* < 0.05; ~, *P* < 0.1; blank, NS.

**Table S3:** Comparisons of the standard deviations (*SD*) of residuals from PC1-based and multivariate (MV) models using datasets both including excluding vascular plant species richness hotspots across the GCFR and SWAFR, across the three spatial scales. Hotspots excluded from each model were those with residuals greater than two standard deviations from the mean for that model. *F*-tests of the ratios of GCFR to SWAFR *SD*s in each case were all significant (*P* < 0.01) except at the DS-scale (c) when species richness hotspots were excluded (denoted by † (*P* = 0.337) and ‡ (*P* = 0.466)).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | *SD* of model residuals | | | | |
|  |  |  | Including hotspots | |  | Excluding hotspots | |
| Scale | Region |  | PC1 | MV |  | PC1 | MV |
| (a) QDS | GCFR |  | 335.17 | 315.52 |  | 233.45 | 222.60 |
|  | SWAFR |  | 247.39 | 230.05 |  | 198.36 | 174.13 |
| (b) HDS | GCFR |  | 607.10 | 540.17 |  | 467.83 | 437.31 |
|  | SWAFR |  | 387.00 | 337.31 |  | 343.84 | 299.52 |
| (c) DS | GCFR |  | 965.29 | 638.38 |  | 665.02 † | 383.79 ‡ |
|  | SWAFR |  | 558.41 | 353.93 |  | 554.68 † | 336.43 ‡ |

# Figures



**Figure S1:** Pairwise correlation coefficients (upper-right panels), scatter plots (lower-left panels) and distributions (diagonal panels) of different forms of (log10-transformed) environmental heterogeneity (QDS-scale) across the GCFR and SWAFR, demonstrating the broad independence of these variables across the study regions. Abbreviations follow that in Tables S1–S3.



**Figure S2:** Pairwise correlation coefficients (upper-right panels), scatter plots (lower-left panels) and distributions (diagonal panels) of different forms of (log10-transformed) environmental heterogeneity (HDS-scale) across the GCFR and SWAFR, demonstrating the broad independence of these variables across the study regions. Abbreviations follow that in Tables S1– S3 and Figure S1.



**Figure S3:** Pairwise correlation coefficients (upper-right panels), scatter plots (lower-left panels) and distributions (diagonal panels) of different forms of (log10-transformed) environmental heterogeneity (DS-scale) across the GCFR and SWAFR, demonstrating the broad independence of these variables across the study regions. Abbreviations follow that in Tables S1–S3 and Figures S1 and S2.



**Figure S4:** Scatter plots of the first (PC1) and second (PC2) axes following PCAs of the nine forms of environmental heterogeneity (log10-transformed and re-scaled) across the GCFR and SWAFR, calculated at the (a) 0.10°×0.10°-, (b) QDS-, (c) HDS- and (d) DS-scales. The percentage of variance in environmental heterogeneity explained by each axis is noted in parentheses in each panel. Arrows for each heterogeneity variable show each variable’s associations with PC1 and PC2. Abbreviations follow that in Tables S1–S3 and Figures S1–S3.

**Figure S5 (next page):** Grid-cells identified as having outstanding vascular plant species richness (i.e. hotspots) across the GCFR and SWAFR, following univariate regressions (Table 1, Figure 4) against the major axis of environmental heterogeneity (PC1) from the PCA (Figure S4). Hotspots were identified as those cells with residual richness greater than two standard deviations from the mean for that model. No hotspots were found for the SWAFR at the DS-scale. Map projection used: WGS84.

A close up of a map

Description automatically generated

**Figure S6 (next page):** Grid-cells identified as having outstanding vascular plant species richness (i.e. hotspots) across the GCFR and SWAFR, following multivariate regressions (Table S2, Figure 5) against the nine axes of environmental heterogeneity. Hotspots were identified as those cells with residual richness greater than two standard deviations from the mean for that model. No hotspots were found for the SWAFR at the DS-scale. Map projection used: WGS84.

**A close up of a map

Description automatically generated**

A close up of a map

Description automatically generated

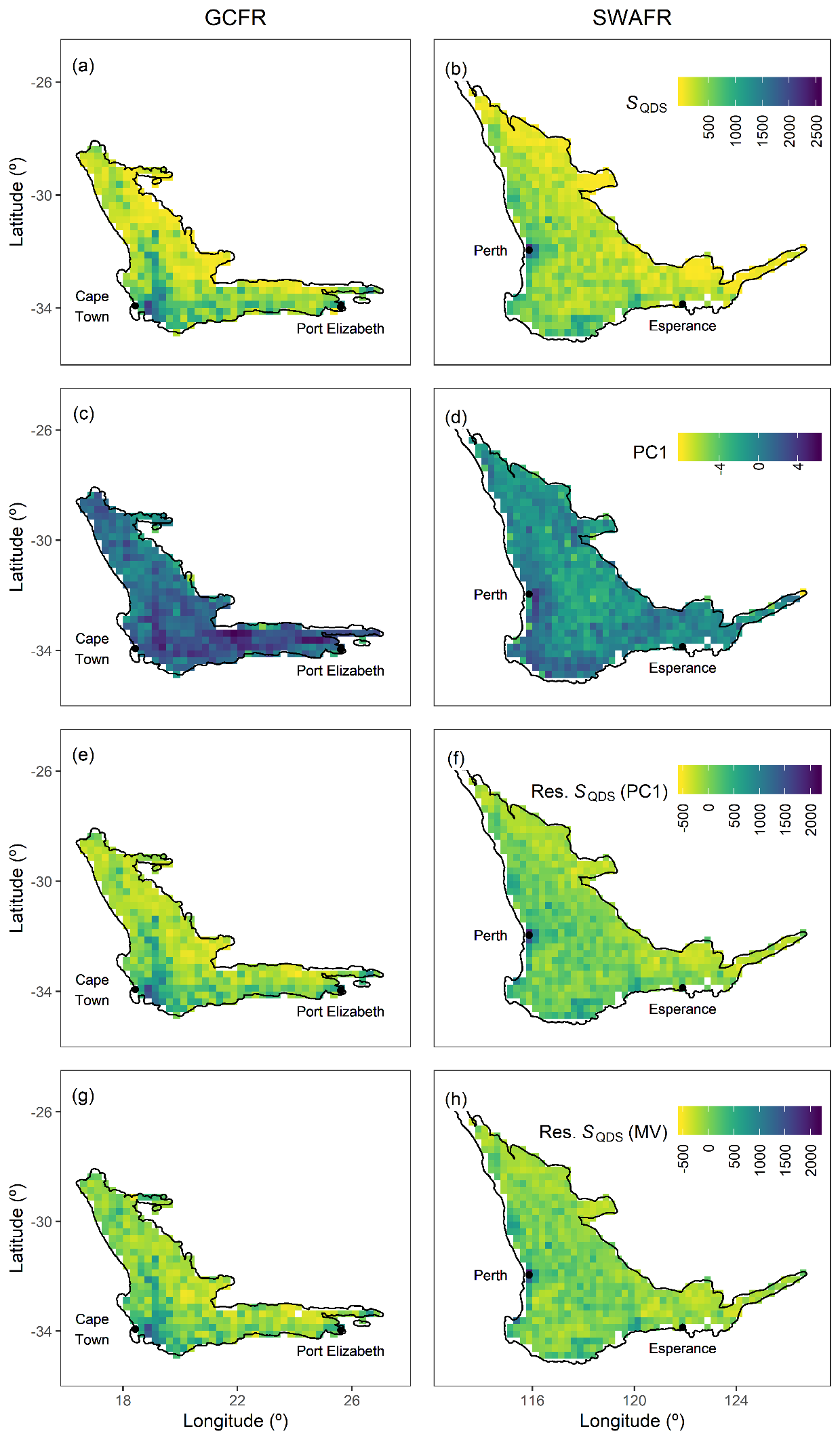
**Figure S7:** Simple linear regressions of vascular plant species richness, excluding species richness hotspots (defined in terms of their residuals from the full PC1-based ANCOVAs; Figure S5), as (a) *S*QDS, (b) *S*HDS and (c) *S*DS against each respective scale’s major axis of environmental heterogeneity (PC1) across the GCFR and SWAFR. These three linear models all have highly significant slopes (*P* < 0.001). For *S*QDS (a) and *S*HDS (b) the separate fits for the GCFR (black) and SWAFR (grey) are presented, following the best fitting model at those scales: a “main effect + region” model (*S* ~ *β*0 + *β*1*X* + *β*2*Region*). The *R*2-values of each model and the variation in environmental heterogeneity explained by PC1 from each of the three PCAs are noted in parentheses in the panel and horizontal axis headings respectively.

A screenshot of a cell phone

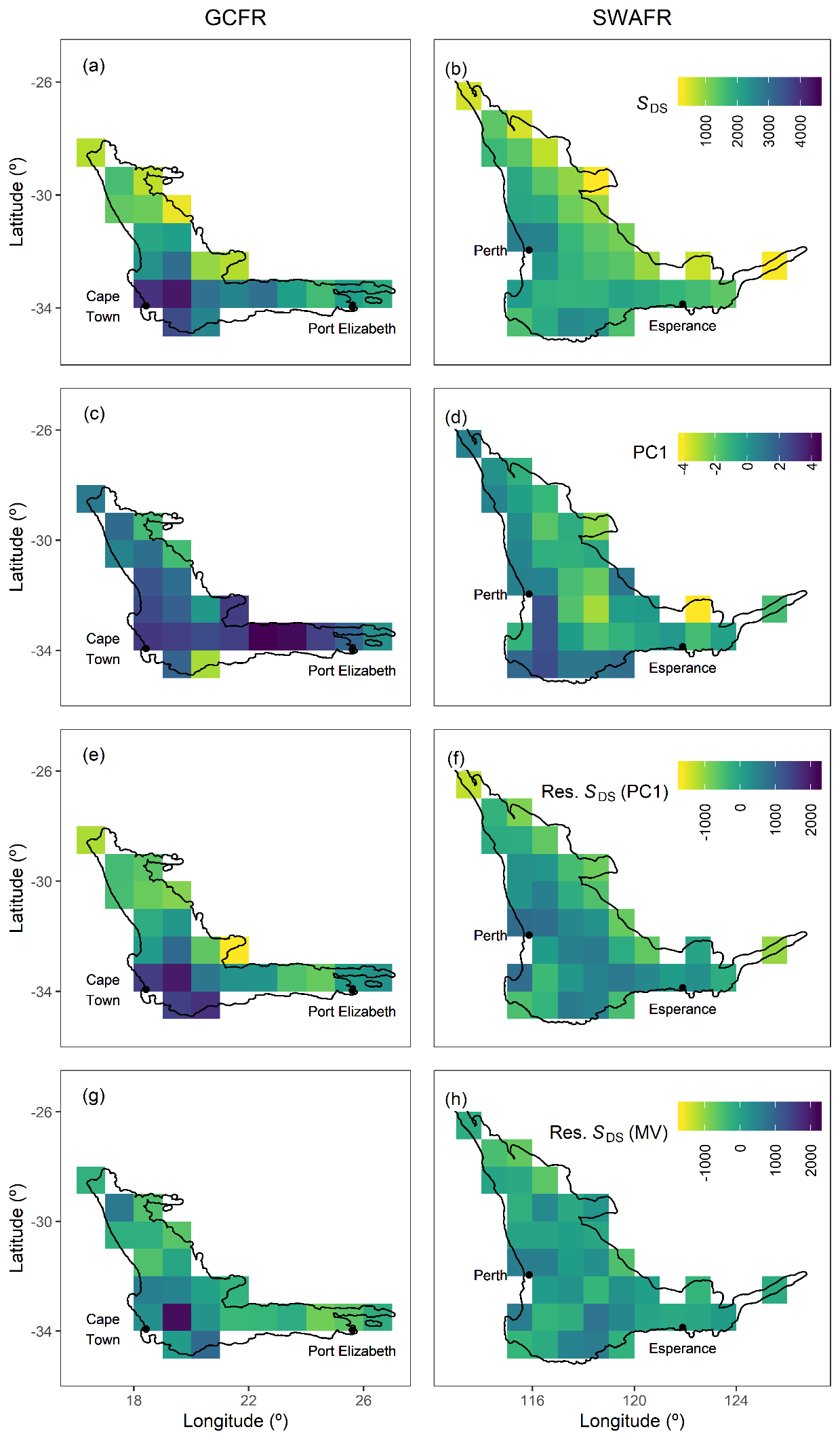
Description automatically generated

**Figure S8:** Slope estimates from multiple linear regressions of vascular plant species richness, excluding species richness hotspots (defined in terms of their residuals from the original multivariate regressions; Figure S6), as (a) *S*QDS, (b) *S*HDS and (c) *S*DS against the various forms of environmental heterogeneity (log10-transformed and re-scaled) across GCFR and SWAFR. Points with error bars denote partial effect estimates and their 95% confidence intervals. Filled and empty points represent effect estimates for the GCFR and SWAFR respectively when region-interaction terms were retained during stepwise model selection, while crosses represent main effects (i.e. no region-interaction term retained). Estimates illustrated in black were significant (*P* < 0.05), while those in grey were not, but still retained during stepwise model selection. The multiple adjusted *R*2-values of each model are noted in parentheses in the panel headings. Abbreviations follow that in Tables S1–S3 and Figures S1–S4.

**Figure S9 (next page):** QDS-scale maps for the GCFR and SWAFR of (a,b) vascular plant species richness, (c,d) the major axis of environmental heterogeneity (PC1) from the PCA of nine forms of environmental heterogeneity (log10-transformed), residuals from regressions of species richness against (e,f) PC1 (Figure 4b) and (g,h) the multivariate (MV) model (Figure 5b). Map projection used: WGS84.



**Figure S10 (next page):** DS-scale maps for the GCFR and SWAFR of (a,b) vascular plant species richness, (c,d) the major axis of environmental heterogeneity (PC1) from the PCA of nine forms of environmental heterogeneity (log10-transformed), residuals from regressions of species richness against (e,f) PC1 (Figure 4b) and (g,h) the multivariate (MV) model (Figure 5b). Map projection used: WGS84.



# References

Boyle, B., Hopkins, N., Lu, Z., Garay, J. A. R., Mozzherin, D., Rees, T., … others. (2013). The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, *14*(1), 16.

Chamberlain, S., Szocs, E., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., … O’Donnell, J. (2016). taxize: Taxonomic information from around the web. R package version 0.7.8. Retrieved from <https://github.com/ropensci/taxize>

Farr, T., Rosen, P., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., & Alsdorf, D. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, 45, 1–33. <https://doi.org/10.1029/2005RG000183>

Funk, C.C., Peterson, P.J., Landsfeld, M., Pedreros, D.H., Verdin, J., Shukla, S., Husak, G., Rowland, J.D., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2, 150066. <https://doi.org/10.1038/sdata.2015.66>

GBIF (24 July 2017a). Global Biodiversity Information Facility (GBIF) Occurrence Download. <https://doi.org/10.15468/dl.n6u6n0>

GBIF (24 July 2017b). Global Biodiversity Information Facility (GBIF) Occurrence Download. <https://doi.org/10.15468/dl.46okua>

Global Names Resolver (GNR). <http://resolver.globalnames.org/>

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagoti?, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017) SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE*, 12, e0169748. <https://doi.org/10.1371/journal.pone.0169748>

Integrated Taxonomic Information System (ITIS). <http://www.itis.gov>

NASA (2017a). MOD13C2 MODIS/Vegetation Indices Monthly L3 Global 0.05Deg CMG (MOD13C2) v006. NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, U.S.A. <https://doi.org/10.5067/MODIS/MOD13C2.006>

NASA (2017b). MOD11C3 MODIS/Land Surface Temperature Emissivity Monthly L3 Global 0.05Deg CMG (MOD11C3) v006. NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, U.S.A. <https://doi.org/10.5067/MODIS/MOD1132.006>

Taxonomic Name Resolution Service (TNRS) v4.0. <http://tnrs.iplantcollaborative.org/>