New regionally modelled soil layers improve prediction of vegetation type relative to that based on global soil models

Michael D. Cramer | Lara M. Wootton | Ruan van Mazijk | George Anthony Verboom

Abstract

**Aim:** High-resolution spatial soil data are crucial to species distribution modelling for fundamental research and conservation planning. Recent globally modelled soil layers (e.g. SoilGrids) have transformed distribution modelling, but may fail to represent regional soil characteristics accurately. We hypothesize that in the Cape biodiversity hotspot of South Africa, the use of global soil layers has led to underestimation of the importance of edaphic factors as determinants of species’ and vegetation distributions. We present a series of new, regionally modelled layers to address this deficiency.

**Location:** Greater Cape Floristic Region (GCFR, South Africa).

**Methods:** We georeferenced edaphic characteristics from literature and other sources and used boosted regression trees (BRT) to associate edaphic characteristics with spatially explicit topographic, climatic, soil texture and biotic variables. Multinomial BRTs were used to predict mapped vegetation types from the collated edaphic and other data.

**Results:** BRTs reliably predicted pH (92% of variance), Na (87%), K (85%), electrical conductivity (81%) and P (73%), but were less accurate for total N (55%) and total C (61%), for which data were sparser. Soil clay and pH values differed markedly in range and in spatial variation from those in SoilGrids. Using our new edaphic layers, we were able to accurately predict spatial distributions of vegetation types within the GCFR (multi-class AUC = 0.96). The multinomial BRT predicted vegetation less well when based on SoilGrids data alone (AUC = 0.84).

**Main conclusions:** The more faithful representation of soil properties in our model is attributable both to its use of ca. 10-fold more samples, and to its regional focus. Our model of edaphic characteristics captures important edaphic variability that is vital for understanding plant and consequently faunal distributions, with wide-ranging conservation implications. Ongoing development of global syntheses of soil data requires more samples, especially in areas with high spatial heterogeneity and extreme edaphic conditions.

**Keywords**

Cape Floristic Region, Harmonized World Soils Database, soil properties, SoilGrids, species distribution models, species richness
Freely available geospatial datasets have transformed biological research capabilities by providing extensive information on environmental conditions (Hampton et al., 2013; Ladeau, Han, Rosi-Marshall, & Weathers, 2017). The biological community has made good use of this information to address a diversity of questions, commonly through the application of species distribution models (SDMs, Elith & Leathwick, 2009). These data have been used to model the distributions of individual species and species assemblages, often in combination with projections under future climate scenarios. The associations between species distributions and environmental variables revealed by such models have been widely used to inform conservation decisions (Hannah et al., 2007; Midgley, Hannah, Millar, Thuiller, & Booth, 2003).

The combination of satellite-derived measures, widespread point sampling of climate data and sophisticated modelling has resulted in many high resolution and reasonably accurate global products for climate (Karger et al., 2017). Global soil layers are, however, much less well-developed, being available only at coarse resolutions and with limited accuracy (Grunwald, Thompson, & Boettinger, 2011). This lack of soil information has been a particular constraint for plant biologists, though many aspects of faunal biology and distribution are also tightly linked to soils through vegetation characteristics. Two data products have attempted to increase the quality of soil information available for biological studies, especially at global and regional scales. The first, the Harmonized World Soils Database (HWSD), provides one of the few globally extensive sources of soils data, and is created from soil sampling which has been “harmonised” to provide a standardised set of variables in standard units (Nachtgeraele, Velthuizen, & Verelst, 2008). This valuable dataset is based on the association of soil physical, chemical and hydrological properties at particular sample sites, with soil mapping units derived from soil maps. The HWSD has over 16,000 soil mapping units (SMU) linked to the harmonized attribute data (FAO & IIASA, 2009). SMU delineations comprise proportions of individual soil types, referred to as Soil Units (SU; Dewitte et al., 2013). The main limitation of the HWSD is the use of these SMUs, which are defined with variable accuracy and across which soil characteristics change due to local variation in geology, topography and other soil-forming processes. In contrast, the SoilGrids 1 km (Hengl et al., 2014) and the SoilGrids 250 m (Hengl et al., 2017) are global raster soil information datasets that attempt to interpolate soil sample properties using a wide range of environmental correlates. For example, the SoilGrids 250 m represents a high-resolution (250 m) interpolation of 150,000 sample points using 158 remotely sensed covariates that include multispectral satellite data, elevation and its derivatives (e.g., slope), climatic variables, landforms and lithology maps (Hengl et al., 2017). These covariates have been associated with the soils data using machine learning algorithms, resulting in representation of 61% of the variance across all soil characteristics. These soils data have been transformative, particularly for ecological research and have been extensively utilized.

The Greater Cape Floristic Region (GCFR) of South Africa (Born, Linder, & Desmet, 2007) is a global biodiversity hotspot accommodating an estimated ca. 11,500 vascular plant species in an area of ca. 200,000 km² (Snijman, 2013). The floristic diversity of this region, which has been attributed both to its long-term climate stability (Cowling, Bradshaw, Colville, & Forest, 2017) and environmental heterogeneity (Cowling et al., 1997; Cramer & Verboom, 2017; Linder, 1991), is organized into six principal vegetation types: forest, Succulent Karoo, Fynbos, Renosterveld and Subtropical Thicket (Bergh, Verboom, Rouget, & Cowling, 2014). These vegetation types have in turn been divided into 189 vegetation units (Mucina & Rutherford, 2006) which associate with a variety of geologies. As the boundaries between these vegetation units are often sharp, despite being located in a similar winter rainfall-dominated, warm-temperate environment, their names commonly reflect their underlying geologies (Power, Verboom, Bond, & Cramer, 2017). For example, 31 vegetation units in the GCFR are designated as Sandstone Fynbos (Mucina & Rutherford, 2006). In the Amazon, similarly, edaphic turnover (assessed from texture and soil classification) plays an important role in determining vegetation community composition, alongside topography and climate (Baldeck, Tupayachi, Sinca, Jaramillo, & Asner, 2016). These authors found that these drivers operate at different scales, with the effect of climate being regional, and edaphic and topographic factors being more important at the landscape level.

The observation that the high biodiversity of the GCFR coincides with high environmental heterogeneity led Cramer and Verboom (2017) to test the relationship between species richness and the heterogeneity of climatic, biotic and edaphic properties. While environmental variables and their heterogeneities together accounted for 95% of biome-level species richness in the region, the inferred role of edaphic properties in determining species richness was relatively minor. This result was surprising given that, although the absence of any association between alpha diversity and soil properties has been noted previously (Cowling & Campbell, 1983), regional (gamma) species richness has been shown to correlate with edaphic heterogeneity (Cowling, 1990). Cramer and Verboom (2017) attributed the low retention of edaphic properties in their species-richness models to a paucity of precise soil data, despite their use of the 1-km resolution SoilGrids.

We hypothesized that inadequate regional soil sampling and global estimation of model parameters limit the utility of globally interpolated soil products of soils in an area such as the GCFR, which is characterized by high spatial turnover in geological and edaphic properties. In such environments, if not generally, locally trained predictions of soil properties are likely to be more accurate (Hengl et al., 2017). We therefore predicted that improved regional soil sampling and the application of models to locally relevant predictor variables would yield significantly improved interpolated soil maps. We also hypothesized that regionally modelled layers of quantitative soil chemical properties would yield better predictions of vegetation distribution than those based on SoilGrids. To test these ideas, we collated edaphic data from published and unpublished sources and
georeferenced the collection sites. We then used a machine learning algorithm, similar to that used for SoilGrids, to interpolate these data spatially. The resulting layers were then used to predict vegetation type distributions, and the predictions compared with those based on SoilGrids 250m. The existence of regionally modelled soil maps, alongside globally modelled maps, is likely to benefit a diversity of researchers.

2 | METHODS

2.1 | Literature and data search and georeferencing

To obtain published data on soil edaphic characteristics, a literature search was conducted within the Scopus database. We also searched the “grey” literature accessible at https://open.uct.ac.za/handle/11427/7909 and http://scholar.sun.ac.za/. Studies were excluded if they dealt with the effects of invasive species on soil properties or were obviously conducted on agricultural land. All of the resulting papers were scanned for usable data. If it was apparent that a paper had used soil information but had not presented the raw data, the corresponding author was emailed to request access. In addition, emails requesting data were sent to individuals who might have worked with soil data in the GCFR, including those in microbiology, environmental and geographical sciences, and biology departments at universities across South Africa, as well as to independent consultants and government bodies. Since response rates to these emails were low, most of the data used are published or from theses (a list of the data sources is found in Appendix S1). The majority of the sourced data were georeferenced, but accuracy varied from a matter of metres to up to 2 km. Where the data were not georeferenced but the authors had provided a description of the locality, we attempted to find it on Google Earth, and used the data if we were certain of the location to an accuracy of greater than 2 km.

The most commonly available soil characteristics were extracted from the compiled data. The characteristics retained included soil pH, electrical conductivity (EC, mS/m), cation exchange capacity (CEC, cmol$^+$ kg$^{-1}$), total C (% w/w), organic C (% w/w), total N (% w/w), extractable P (mg/kg), extractable K (cmol$^+$ kg$^{-1}$) and extractable Na (cmol$^+$ kg$^{-1}$). Although soil pH was variably measured either in water (pH$_{\text{H}_2\text{O}}$) or in 1 M KCl (pH$_{\text{KCl}}$), the presence of a strong linear relationship (pH$_{\text{KCl}}$ = 1.11·pH$_{\text{H}_2\text{O}}$-1.47; $r^2$ = 0.97) between pH$_{\text{H}_2\text{O}}$ and pH$_{\text{KCl}}$ in those samples for which both measures were available enabled us to convert all pH$_{\text{H}_2\text{O}}$ values to pH$_{\text{KCl}}$. Extractable P values were either derived from Olsen (Olsen, Cole, Watanabe, & Dean, 1954) or Bray II P (Bray & Kurtz, 1945) and were assumed to be equivalent. Extractable cations were extracted with ammonium acetate.

Using their geographic coordinates, the soil values were associated with a vegetation classification. Each South African biome, which is broadly defined by climatic conditions and dominant plant growth-forms (Rutherford & Westfall, 1986), has been classified into vegetation units (Mucina & Rutherford, 2006). The vegetation classification (Vegmap 2012 beta; http://bgis.sanbi.org), which is updated from the vegetation classification of Mucina and Rutherford (2006), was recoded so that Renosterveld and Strandveld vegetation categories, commonly considered as part of the Fynbos Biome, were listed separately in accord with the fact that the generic turnover between Fynbos, Renosterveld and Strandveld is almost complete (Bergh et al., 2014). These vegetation groupings are subsequently referred to as “vegetation types”, which are distinct from the “vegetation units” in the classification of Mucina and Rutherford (2006). The edaphic properties were associated with the vegetation types in which the points were located and subjected to one-way analysis of variance (factor vegetation type) to determine the significance of differences between the vegetation units. All statistical analyses were conducted in R (R Core Team, 2018).

2.2 | Environmental data

The floristic composition of the vegetation at each soil sampling point was quantified using the first three axes used by Bergh et al. (2014) to represent patterns of genus-level floristic similarity between the vegetation units making up the Succulent Karoo and Fynbos Biomes (sensu Mucina & Rutherford, 2006). These axes, which are based on the “important species” lists provided by Mucina and Rutherford (2006), were extracted by Bergh et al. (2014) using classification analysis. For each soil sampling point, geographical coordinate data were first used to identify the vegetation unit present, and this then converted to floristic scores using the axes of Bergh et al. (2014).

Climate data were obtained from Fick and Hijmans (2017) for the period 1970–2000 at ca. 1 km$^2$ resolution. Variables included are mean annual temperature (MAT), mean annual precipitation (MAP) and 17 other derived bioclimatic variables. Additionally, we included solar radiation, wind speed and vapour pressure from this data source. Potential evaporation (PET, mm) was obtained from Zomer, Trabucco, Bossio, Straaten, and Verchot (2008). Normalized difference vegetation index (NDVI) data were obtained for southern Africa from eMODIS TERRA (US Geological Survey Earth Resources Observation and Science Center) which is corrected for molecular scattering, ozone absorption and aerosols. The data from a time series (2001–2010) with smoothing at a 250-m spatial resolution (1999, & Marko, 11999) were averaged to obtain monthly and annual average values.

For comparison purposes, soils data were extracted from the African Soil Profiles (AFSP) database which is a combination of AFSP with >17,000 georeferenced legacy soil profile records and AfSIS Sentinel Site database with >9,000 sampling locations (Hengl, 2017). Soils data were also obtained from SoilGrids at 250 m resolution (Hengl et al., 2017). Variables included in the analysis were pH (KCl extract), coarse-, sand-, silt- and clay-fractions (% w/w), organic carbon (% w/w), cation exchange capacity (CEC, cmol$^+$ kg$^{-1}$) and bulk density (kg/L). These data for six soil depths to 2 m were averaged by depth-weighted averaging (i.e. averaged by weighting values for each depth interval; Hengl et al., 2017). Elevation data from the Shuttle Radar Topography Mission (SRTM) at 30-m horizontal and 16-m vertical resolution (https://earthexplorer.usgs.gov, accessed
Nov 2017) were used to calculate landscape slope and aspect using the terrain function in the "raster" library (Hijmans, 2016) in R. The aspect data were recoded to eight cardinal directions.

Fire frequency was determined between 2000 and 2014 from the MODIS burned area product (MODIS Burned Area Product 5.1, 2014; modis-fire.umd.edu). The confidence of detection band was used to reclassify pixels over agricultural areas as missing data. Pixels that lacked valid data for >7 months were also reclassified as missing data. Because of the lack of the satellite data in June 2001, each month over the 14 years was averaged and the monthly averages were then averaged to acquire the mean annual fire frequency.

Geological data (chrono-stratigraphic) were obtained from the Council for Geosciences (http://www.geoscience.org.za). In this, we consolidated the parent geology fields in this data set (i.e. combining "parent1" and "parent") and also used the chronology description "Quaternary" to differentiate Quaternary sands. Where this consolidated parent field was undefined, we used the lithostratigraphic information (" lithstrat" field). As links to soil properties, these geological descriptions are potentially misleading because the geological descriptions pertain to the underlying geology, rather than the overlying soil origins which might transported soils (e.g. colluvium, alluvialite). The land types of South Africa (Land Type Survey Staff, 1972-2006) were used to obtain the soil pattern code of each sample point. The land types are classified as areas with relatively uniform terrain form, soils and climate. These data also include soil depth, topsoil clay (% w/w) and the occurrence of rocky outcrops. Land Cover Thematic Data for 2013-2014, which differentiates 72 classes of land use, were obtained from the South African Department of Environmental Affairs (https://egis.environment.gov.za). These data differentiate natural vegetation, cultivated land, urban areas and other transformations.

2.3 | Boosted regression tree analysis

Boosted regression tree (BRT) model construction was performed following Elith, Leathwick, and Hastie (2008) using the "dismo" library (Hijmans, Phillips, Leathwick, & Elith, 2017) in R. The variables that were initially included in the model prior to simplification were the climate variables, the topographic variables, NDVI, fire, vegetation units and types, geology and land types together with clay, soil depth and presence of rock. We removed categorical predictors from our models when we realised that the BRT modelling algorithm generates spurious associations with highly multinomial categorical variables, especially when some categories contain limited or no data. For this reason, the final model was generated using the floristic scores but not geological categorisations, vegetation units or vegetation types.

A Gaussian BRT model was used to predict individual soil characteristics for which there was sufficient data available in the georeferenced data. The individual soil characteristics that were included were pH and the log of total C, total N, extractable K, extractable Na, extractable P, and electrical conductivity (EC). The log of these values was used to accommodate the range of values in the model. Following optimisation with variable tree complexities (1-5) and learning rates (0.01, 0.005, 0.001, 0.0005, 0.0001), the tree complexity was set to 5 and learning rate to 0.001 with fivefold cross-validation and a bagging fraction of 0.7. These parameters were supplied to the gbm.step function in the "dismo" library in order to build the BRT models with optimisation by cross-validation, following Elith et al. (2008). The initial models were simplified, and the importance of predictor variables in determining the response variables was ranked. The models were generated both with and without NDVI as a predictor variable (see below). For each of the fitted models, the pseudo-$r^2$ was calculated as 1−(residual cross-validation deviance/total deviance). To ensure that the BRT results were not a product of chance alone (Calude & Longo, 2017), the response variables were reassigned randomly to environmental variables and the BRT models generated as before. These randomized models all failed to reach a solution, verifying that the model structure is not spurious. Following model development, the model was evaluated to determine the degree of deviance explained and the predictions plotted against the observed data to determine the correlation between observed and predicted values. The models were used to predict the soil characteristics using the predict function in the "dismo" library against rasters of the environmental data used to build the models. In order to fit the models, some response variables with large ranges of values were logged because of the edaphic heterogeneity of the GCFR. Because log transformation results in systemic overprediction at low-observed values and overprediction at high-observed values, the spatially explicit predictions (only) were corrected using the linear regression of predicted on observed values (predicted value = intercept/slope). We used the "terrain" function in the "raster" library (Hijmans, 2016) for R to calculate "roughness" in relation to the eight neighbouring pixels to gauge heterogeneity of soil pH from the SoilGrids and our modelled data. The "roughness" metric yields the difference between the maximum and minimum values between a focal pixel and its eight neighbour pixels.

The modelled soil properties (pH, total C, total N, ext. K, ext. Na, ext. P, EC) together with variables extracted from the land types (clay, soil depth and the presence of rock) were extracted in a regular grid and incorporated into a multinomial BRT model to predict vegetation types produced using the "gbm" library (Ridgeway, 2017). The model was run through 2,200 iterations (established as optimal using fivefold cross-validation) with shrinkage = 0.01, interaction depth = 5, bag fraction = 0.5 training fraction = 0.75. In addition, soil data derived without the inclusion of NDVI and genus-level floristic similarity indices were also used to predict vegetation types. This was done to ensure that there was no potential circularity as a consequence of vegetation characteristics (i.e., NDVI and genus-level floristic similarity indices) first being used to predict soil, and soils then being used to predict vegetation. The relative influence of the predictor variables was summarized, and the "pROC" library (Robin et al., 2011) used to calculate the multi-class area under the curve (AUC). The vegetation type corresponding to the maximum probability for each sample point was extracted and mapped. Similarly, the modelled soil properties were combined with other climatic,
topographic, NDVI and fire data (see Section 2.2) to allow combined prediction of vegetation types using a multinomial BRT. We also predicted the vegetation types with all non-edaphic variables and/or edaphic data extracted from SoilGrids.

3 | RESULTS

3.1 | Soil sampling and comparison of data sources

The density of sampling varied between the different soil characteristics (Figure 1). In addition, the density of sampling in our compilation is much greater in the coastal areas than in the interior, whereas the AFSP data cover a greater geographic range within the GCFR. The different sources of data, however, provided quite distinct values. For example, the amount of clay from AFSP was generally higher than that reported for the land types data (Figure 1, see Appendix S1 in Fig. S1.1). Apart from being higher overall, the AFSP clay proportions are high along the south coast of the GCFR, increasing towards the north-west. From the land types polygon data, which are based on soil classification data, there is a tendency for clay to be higher in the north-west. Overall, the median proportion of clay was 8.2% (range: 4.6–20.0) versus 13.2% (range: 0–37.5) in the SoilGrids data (Figure 2). Similarly, the soil pH was lower in our compilation than that reported in the AFSP (Figure 1) or SoilGrids datasets (see Appendix S1: Fig. S1.2). Despite being significantly correlated ($r^2 = 0.289$), the observed GCFR values used for our model development departed strongly from a 1:1 relationship with the SoilGrids pH data. The lack of correspondence between these datasets extends to organic C which also departed strongly from a 1:1 relationship with the SoilGrids organic C data (Fig. S1.2).

3.2 | Variation in soil characteristics with vegetation of the GCFR

The vegetation types in the GCFR for which we had adequate data were forest, Fynbos, Renosterveld, Strandveld and Succulent Karoo.
Although the MATs of these vegetation types were similar, Succulent Karoo is the most arid (Table S1). Strandveld occurs close to the coast and at low elevations. Succulent Karoo and forest had the lowest and highest NDVI values, respectively. Fires were common in forest and Fynbos areas, but very rare in Renosterveld and Strandveld and absent in Succulent Karoo. The occurrence of fires in forest (see Appendix S1: Table S1) is at odds with the fact that fires generally singe the edges of forests, but seldom burn through the forests (Kraaij & van Wilgen, 2014). This anomaly is due to the fact that forest patches are small, typically below the spatial resolution of the fire data used in this analysis (ca. 1 km²), and commonly embedded in small, fire-protected sites within fire-prone Fynbos vegetation.

Since soil heterogeneity over even small spatial distances is common, it is unsurprising that there was considerable variability in many of the measures of soil characteristics, even within vegetation units (Figure 3). This is perhaps best illustrated by pH which was most densely sampled (Figure 1). Despite substantial variation in soil nutritional attributes between vegetation types, the individual soil variables did not vary in a consistent manner except in respect of Fynbos which had relatively low EC, pH, total N and extractable P, K and Na. For example, although Strandveld has high EC, pH and total N and Na relative to Fynbos, extractable P and K are similar, and organic C is lower than that of Fynbos. This weak association with the vegetation types is likely because each of latter includes multiple sub-types (i.e. vegetation units in Mucina & Rutherford, 2006) which occur on different soils. For example, Fynbos ranges from acidic sandstone Fynbos to limestone Fynbos. As a consequence, modelling efforts need to be able accommodate this variability, and also the potential interactions between predictors.

3.3 | Boosted regression models

Use of machine learning algorithms such as BRTs for modelling spatial variation in response variables requires the association of these variables with predictor variables. Within the GCFR, several environmental variables show covariation (see Appendix S1: Fig. S1.3) which is strongly associated with topography (e.g., MAT, MAP), distance from the coast (e.g., wind) and latitude (e.g., solar radiation). Sufficient soil sample data were only available to enable the construction of reliable models for pH, EC, total C, total N and extractable P, K and Na. Predictor variables contributing significantly to these models were retained following model simplification (see Appendix S1: Fig. S1.4). Although elevation (retained in two of seven models) is collinear with several environmental predictors (e.g. MAT and MAP), we retained it in our models because our focus was on developing a model for predicting soil characteristics, rather than explaining their environmental dependencies. The BRT models identified a diverse collection of most influential predictors across the modelled response variables. The only variable retained in all predictions of soil chemical properties was soil clay concentration, derived from the spatially explicit land type data, while soil depth was influential in three of the seven models. NDVI was retained in five models, indicating a strong association between soil properties and vegetation greenness. Despite this, omission of NDVI and genus-level floristic similarity indices had little consequence for the predicted soil properties, reducing pseudo $r^2$ values by less than 2% (data not shown). The model for pH predicted 92.4% of the observed variance with a slope of 0.89 between observed and predicted pH (Figure 4). Although predictions of other variables were all weaker than for pH, the slopes were all 0.45 or higher and the coefficients of determination were greater than 0.55.

We compared the predicted pH from our model with that from SoilGrids (Hengl et al., 2017). The modelled GCFR data capture a much greater spatial variability in soil pH than did the SoilGrids (Figure 5). In particular, the higher pH in coastal areas with a strong marine influence was not accurately represented in SoilGrids. Apart from this, the broad trend of lower pH sites inland of the coast and higher pH in the more arid interior areas was captured by SoilGrids.
A comparison of the SoilGrids pH with the predicted pH from the modelled GCFR compilation at sampled points in the GCFR compilation had a slope of 1.5 and an intercept of -3.4 (see Appendix S1: Fig. S1.5). This indicates that soil pH in the SoilGrids data set is ca. 3.4 pH units higher in the GCFR than it is in our model, and that the correspondence between the two models is relatively weak ($r^2 = 0.36$). In addition, the modelled GCFR compilation results show higher spatial "roughness" (i.e. variation between neighbouring pixels) in soil pH than do the SoilGrids data (Figure 5). This shows that our model captured much more of the fine-scale (ca. 1 km$^2$) variability across the region.

The BRT predictions of soil variables other than pH were less precise (Figure 4) and this needs to be considered when interpreting the patterns (Figure 6). The weakest relationships were with total...
C (61%) and total N (55%). This could mean that there are sources of variation for these response variables that were not captured in the models (e.g., biotic variables or analytical differences), though even these soil characteristics are reasonably well predicted. Soil EC was highest along the low-elevation coastal platform and Succulent Karoo areas (Appendix S1: Fig. S1.3). Although there are no areas with very high extractable P (90th percentile = 24.9 mg/kg), the lowest P values were along the south coast, and mountainous areas.

Total N was lowest along the south and west coast and extending into the mountain ranges with soil C following a similar pattern. Na and K were both lowest along the south coast and in the mountain ranges.

The multi-class AUC of the receiver operating characteristic (ROC) for the prediction of vegetation type from our modelled edaphic data was 0.96, indicating a reasonably accurate prediction of vegetation types on the basis of soil properties alone. When
vegetation was predicted on the basis of soil characteristics modelled without the inclusion of NDVI of genus-level floristic similarity indices, the multi-class AUC was also 0.96 (see Appendix S1: Fig. S1.6) and the resultant vegetation map was indistinguishable from that produced with inclusion of NDVI (Figure 7). The model correctly identified all forest points, 93% of Fynbos, 86% Renosterveld, 98% Strandveld and 83% Succulent Karoo pixels. Thus, the modelled distribution of vegetation closely resembled the mapped vegetation, except for over-prediction of Strandveld vegetation along the west coast (Figure 7), accounting for relatively unsuccessful prediction of Succulent Karoo. To some extent, over-prediction of Strandveld might be due to the diversity of this vegetation and some local debate about the association of some vegetation with Strandveld. Extractable K was the strongest predictor of vegetation type in the multinomial model followed by total C and clay (see Appendix S1: Fig. S1.7). The predicted probability of each vegetation type is differently influenced by each of the edaphic variables used as predictors (see Appendix S1: Fig. S1.8). For example, Fynbos probability decreased with increasing pH, extractable P, K, Na and clay. These relationships are, however, difficult to interpret because they represent the overall dependence of multiple component vegetation units (e.g., multiple Fynbos types lumped under the collective category Fynbos) and the vegetation type probabilities are all inter-dependent. While the multi-class AUC of the edaphic model was relatively high, a comparable model including only non-edaphic variables (i.e., including climatic, topographic, fire and NDVI variables) had a multi-class AUC of 0.84, and the model based on SoilGrids data alone also had a multi-class AUC of 0.84. The main failure of the non-edaphic model was in under-predicting Succulent Karoo and over-predicting Fynbos and Strandveld. The model based only on SoilGrids tended to produce graded boundaries between vegetation types rather than to represent the fact that the real vegetation boundaries are relatively sharp. When all available variables (i.e., climatic, GCFR compilation derived edaphic, topographic, fire and NDVI) were considered, however, six of the top ten predictors were edaphic variables (see Appendix S1: Fig. S1.9), and the model had a multi-class AUC of 0.94.
DISCUSSION

As anticipated by the authors of SoilGrids (Hengl et al., 2017), our regional model of soil properties yielded a more faithful representation of the soil properties of the GCFR than their global map. The main reason for this is that our analyses made use of many more sample points. The SoilGrids model depends on data made available through national soil agencies, with only 209 sites from the GCFR region included, many of which had incomplete data. The GCFR is a highly diverse region with a high degree of environmental turnover, largely driven by topographic variability and oceanic climate buffering along the coastal areas that decline towards the interior (Cramer & Verboom, 2017). This regional heterogeneity makes sample density critical for successfully modelling the highly heterogenous edaphic characteristics of the region (Cramer, West, Power, Skelton, & Stock, 2014a).

A second possible reason for the better quality of the local-scale model is that the South African soils agency data have an agricultural

![Variation in soil chemical properties predicted from the BRT model. The colour scales for the maps represent logs of chemical properties, apart from pH. The coast and the buffered extent of the modelled region that includes the Greater Cape Floristic Region (GCFR) are shown for reference. Map projection is Plate Carrée (Geographic)](image-url)
affiliation (Agricultural Research Council, Institute for Soil, Climate and Water, Pretoria) and is consequently biased towards agricultural areas. This is borne out by the fact that 48.3% of AFSP compared with 9.4% of GCFR compilation sites were in sites identified in the Land Use map as “disturbed” (i.e. agriculture, forestry or urban, see Appendix S1: Fig. S1.10). Some GCFR sampling points identified as “disturbed” might have been on relatively undisturbed road verges adjacent to agricultural fields. Within South Africa, agricultural areas (e.g., wheat farming) are commonly on the more fertile soils (e.g., Renosterveld; Newton & Knight, 2005), with the nutrient-depauperate soils (e.g., those associated with Fynbos) being largely avoided agriculturally (except for indigenous crops of the Proteaceae and Leguminosae). What is unclear is the extent to which the systematic bias towards more fertile soils observed in the regional SoilGrids data set is globally pertinent. Beyond the fact that agricultural practices in the GCFCR are concentrated on soils that are naturally more fertile, it is likely that soils in agricultural areas have been enriched or contaminated with fertilizers, thereby introducing error in the estimation of natural soil properties.

A third possible reason for the improved quality of the local-scale model is that parameter estimation in any global model is inevitably influenced by the characteristics of soils whose values lie outside the range of values observed locally, and is sensitive to environmental predictors which pertain globally but might be unimportant locally. For example, the GCFR region is a predominantly winter rainfall region (see Appendix S1: Fig. S1.11) in which soil development has been influenced by the high leaching potential of winter rains (Cramer & Hoffman, 2015). However, since mediterranean-climate ecosystems comprise only 5% of the global land surface (Belda, Holtanová, Halenka, & Kalvová, 2014), the nuanced effect of a winter rainfall regime is unlikely to be effectively captured in a global synthesis. Furthermore, the GCFR is one of the most nutrient-depauperate...
landscapes in the world due to both a sandstone geology which is regionally but not globally prevalent (Bradshaw & Cowling, 2014) and a long history without glaciation (Hopper, 2009). Thus, the GCFR is particularly difficult to include in a global soil modelling exercise, although similar concerns are likely to apply elsewhere. For example, Kreft and Jetz (2007) presented a model which accounted for ca. 70% of global variation in vascular plant species richness but underestimated that of the Cape Floristic Region of South Africa by half, possibly because they did not consider the consequences of long-term climate stability in the region (Cowling et al., 2017) and because the global model did not adequately capture the heterogeneity of the region (Cramer & Verboom, 2017).

We modelled different soil properties from many of those included in the SoilGrids data set. The reasons for the differences between the properties modelled by SoilsGrids and our model are also informative. Where the latter included soil pedological classifications, water-holding capacity, texture-related variables, pH, CEC and organic C, we modelled pH, EC, total N, total C, extractable P, K and Na. Much of the data analysed were derived from regional ecological studies on naturally occurring plants, in which the soils were analysed using soil science testing methods commonly used for testing agricultural soils. The reported variables are therefore those of interest to physiologists and ecologists in the region. These soil properties represent the chemical and textural components that are thought to determine diverse aspects of the ecology of the flora and fauna of this region. For example, many regional soils have low total N concentrations, averaging 973 mg/kg but with a high degree of variance (214–5,973 mg/kg; Stock & Verboom, 2012), and slow rates of N mineralization. Nitrogen therefore represents an important and highly variable determinant of vegetation characteristics in the GCFR. There is also evidence that P is what has been referred to as “the ultimate limiting nutrient” (sensu Vitousek, Porder, Houlton, & Chadwick, 2010; i.e. P addition transforms ecosystems) in this region (Cramer, 2010) with Cape plants having amongst the highest foliar N:P ratios globally (Stock & Verboom, 2012). The region also has very low clay concentrations, with the median clay concentration being 8.9%, and 25% of land types having soils with < 4.5% clay. This paucity of clay in Cape soils is probably a consequence of the low feldspar and mica levels in Cape sandstones, the limited clay in these soils being instead derived from terrigenous sources of dust (Soderberg & Compton, 2007). The specific edaphic factors that are regionally ecologically important need to be considered when modelling vegetation or faunal distributions within a region.

Compilation of regional edaphic models with higher sampling density is clearly an important step, but also introduces several problems, especially when reliant on pre-existing data. The variables available may be those of local interest or because of historical selection of particular analytical methods. For example, in our analysis we had Olsen or Bray II measures of available P, but there are many other standards for extractable P. Another example is the measure of pH which is commonly with KCl, CaCl₂ or water as extractants, although we were able to cross-correlate between KCl and water extracts. All of these variabilities make compilation of “harmonised” regional datasets difficult and limit the ability to generalize the models to broader areas (e.g. across national boundaries where analytical methodologies may differ). The key to improving global and regional models, we contend, is therefore: (a) standardization of analytical methodology; (b) collection of soil data from a broad range of untransformed habitats; and (c) complete characterization of soil physical and chemical attributes relevant to plant ecology.

Despite their generally leached nature (Figure 3; Cramer & Hofman, 2015), Cape Fynbos soils show substantial nutritional variability between vegetation types. The variability of Fynbos soils is exemplified by the diversity of vegetation units comprising Fynbos (Mucina & Rutherford, 2006), of which there are 81 in the GCFR. The same is true for other vegetation types, but to a smaller extent. For example, there are 62 vegetation units in Succulent Karoo, 28 in Renosterveld, 12 in Strandveld and two in forest. Since the vegetation types are categorical, spatial or floristic mis-classifications can contribute to variability. For example, in the GCFR, Strandveld is confined to dune sands, but has also been mapped as occurring on acidic aeolianites (Mucina & Rutherford, 2006). The high degree of variability within vegetation types is, however, usually less than that between vegetation types, making it possible to differentiate vegetation types on the basis of soil characteristics. The ability to predict vegetation types accurately from the multinomial BRT suggests that the modelled edaphic variables capture important aspects of environmental heterogeneity correlated with vegetation characteristics. Most importantly, however, the edaphic differences between vegetation types are highly variable in stoichiometry. For example, low extractable P in Strandveld soils is not necessarily associated with low total N. This means that use of the concept of “soil fertility”, commonly extracted as an axis in principal components analyses (Coetsee, Bond, & Wigley, 2015; Verboom, Moore, Hoffmann, & Cramer, 2012), is difficult to interpret and may hide much important detail. Thus, while impractical in some situations, we argue that the individual chemical characteristics of soils should be treated as independent wherever possible, and modelling efforts must be able to accommodate the interactions between variables that determine vegetation characteristics.

In the face of rapid global change, forecasting species distributions are central to conservation (Sala et al., 2000). SDMs contribute to conservation in diverse ways including supporting land management, weed, pest and alien species risk assessment and assessing impacts of climate change on biota (Franklin, 2013). The availability of edaphic information is important for a variety of different applications, as evidenced by the fact that SoilGrids models (Hengl et al., 2014, 2017) are both already well cited. For example, floral and faunal SDMs have been produced largely without the use of soil characteristics, because spatially explicit soil data have not been available at scales that are relevant to species distributions (Yates et al., 2010). Particularly, when combined with predictions of the effects of climate change, these SDM models are relatively poorly constrained. The inclusion of SoilGrid data, however, significantly restricts modelled species distributions and the ability of
species ranges to adjust to climate change (Figueiredo et al., 2018). While soil variables have been previously thought to operate at landscape and local scales and climate at larger continental scales, there is evidence to show that soils are important across all scales (Eiserhardt, Svenning, Kissling, & Balslev, 2011; Yates et al., 2010). Although recent authors (Figueiredo et al., 2018) have relied on SoilGrids, these data are unlikely to be regionally accurate (Hengl et al., 2017). A consequence of inaccuracies in available soil data sets is a general failure to identify soil properties as an important determinant of biological pattern. For example, despite the clear importance of soils in limiting the distributions of individual species in South Africa and in influencing regional scale (i.e. gamma) species richness (Cowling, 1990; Richards, Cowling, & Stock, 1997), Cramer and Verboom (2017) failed to identify a strong role for soil attributes or edaphic heterogeneity in determining plant species richness patterns in South Africa despite making use of SoilGrids (1 km resolution, Hengl et al., 2014). This may reflect the failure of SoilGrids to capture the fine-scale variations in edaphic properties and perhaps some of the regionally important variables (e.g. N and P) that determine species richness in the region.

The prediction of the vegetation based on edaphic variables alone does not imply that climate is not important, since some vegetation types were accurately represented by models based on climate alone. Furthermore, soils are the combined product of climate, biota, topographic relief, parent geology and soil age (Jenny, 1941). There is thus a feedback between soil and vegetation characteristics. For example, small forest patches in the GCFR commonly occur on the same geological parent material as neighbouring Fynbos, but the forest soils have higher pH, total N, P, K, Ca, Mg, and organic matter (Cramer et al., 2019). These differences are attributable to the fire-susceptibility of Fynbos, which results in nutrient losses, and the limited flammability of forests which fosters nutrient accumulation. Therefore, while soils may be used to predict vegetation, the vegetation also determines the soil characteristics.

5 | CONCLUSION

The vegetation types of the GCFR, a global-plant biodiversity hotspot, are strongly associated with variation in edaphic characteristics. The association of faunal biodiversity with plant biodiversity (Colville et al., 2014) extends the influence of edaphic characteristics to faunal distributions. This makes it imperative that SDMs and other modelling activities include edaphic characteristics in order to produce accurate maps and/or predictions. Although global soil maps are an invaluable resource, users need to heed the warnings from the authors of these data compilations, because the syntheses are no better than the data used to generate them. The high spatial resolution of these maps (currently 250 m) is seductive for users who might be tempted to assume that these maps are highly accurate. This may be especially problematic in areas such as the GCFR which are unusual in a global context (e.g. Mediterranean climate), show strong edaphic heterogeneity, and whose soils have been sparsely sampled. The maps of edaphic variables that we have produced for the GCFR (available in raster format in the supplementary information) probably represent a more reliable depiction of some aspects of soil characteristics in the region than those in current use. Nevertheless, this represents only an initial step towards a more extensive compilation of data, that should be made freely available.

ACKNOWLEDGMENTS

We are grateful for the ‘land types’ data supplied by the ARC-Institute for Soil, Climate and Water (South Africa) and to others who contributed data. We are also grateful to Lindsey Gilson for arranging funding from the Faculty of Science (University of Cape Town). Strategic Impact Areas call under the ‘Biodiversity and Environmental Change in the Cape Floristic Region’ programme to support data collation.

DATA AVAILABILITY STATEMENT

All point sample data and predicted soil GIS layers (rasters) are available from the Dryad data repository. DOI: https://doi.org/10.5061/dryad.37qc017.

REFERENCES

Baldeck, C. A., Tupayachi, R., Sinca, F., Jaramillo, N., & Asner, G. P. (2016). Environmental drivers of tree community turnover in western Amazonian forests. *Ecography*, 39, 1089–1099. https://doi.org/10.1111/ecog.01575

Belda, M., Holtanová, E., Halenka, T., & Kalvová, J. (2014). Climate classification revisited: From Köppen to Trewartha. *Climate Research*, 59, 1–13. https://doi.org/10.3390/climresearch.59.1-13

Bergh, N. G., Verboom, G. A., Rouget, M., & Cowling, R. M. (2014). Vegetation types of the Greater Cape Floristic Region. In N. Allsopp, J. F. Colville, G. A. Verboom, & R. M. Cowling (Eds.), Fynbos: ecology, evolution and conservation of a megadiverse region (pp. 1–25). Oxford, UK: Oxford University Press.

Born, J., Linder, H. P., & Desmet, P. (2007). The Greater Cape Floristic Region. *Journal of Biogeography*, 34, 147–162.

Bradshaw, P. L., & Cowling, R. M. (2014). Landscapes, rock types, and climate of the Greater Cape Floristic Region. In N. Allsopp, J. F. Colville, G. A. Verboom, & R. M. Cowling (Eds.), Fynbos: ecology, evolution and conservation of a megadiverse region (pp. 26–46). Oxford, UK: Oxford University Press.

Bray, R. H., & Kurtz, L. T. (1945). Determination of total, organic and available forms of phosphorus in soils. *Soil Science*, 59, 39–45. https://doi.org/10.1097/00010694-194501000-00006

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of Science*, 22, 595–612. https://doi.org/10.1007/s10699-016-9489-4
Nachtsgaele, F., van Velthuizen, H., & Verelst, L. (2008). The harmonized world soil database. version 1.0. Fao/lisia/Isric/lss-Cas/Jrc, Version, 1, 34–37.

Newton, I. P., & Knight, R. (2005). The use of a 60-year series of aerial photographs to assess local agricultural transformations of West Coast Renosterveld, an endangered South African vegetation type. South African Geographical Journal, 87, 18–27. https://doi.org/10.1080/03736245.2005.9713822

Olsen, S. R., Cole, C. V., Watanabe, F. S., & Dean, L. A. (1954). Estimation of available phosphorus in soils by extraction with sodium bicarbonate. USDA Circular, 939.

Power, S. C., Verboom, G. A., Bond, W. J., & Cramer, M. D. (2017). Environmental correlates of biome-level floristic turnover in South Africa. Journal of Biogeography, 44, 1745–1757. https://doi.org/10.1111/jbi.12971

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Richards, M., Cowling, R., & Stock, W. (1997). Soil nutrient dynamics and community boundaries in the Fynbos vegetation of South Africa. Plant Ecology, 130, 143–153.

Ridgeway, G. (2017). gbm: Generalized Boosted Regression Models. R package. https://cran.r-project.org/package=gbm

Robinson, V., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). PROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, 77. https://doi.org/10.1186/1471-2105-12-77

Rutherford, M. C., & Westfall, R. H. (1986). Biomes of southern Africa – an objective classification. Memoirs of the Botanical Survey of South Africa, 54, 1–98.

Sala, O. E., Chapin, F. S., Armento, J. J., Berlow, E., Bloomfield, J., Dirzo, R., ... Wall, D. H. (2000). Global biodiversity scenarios for the year 2100. Science, 287, 1770–1774.

Snijman, D. A. (2013). Plants of the great cape floristic region. 2: The extra cape flora. Pretoria, South Africa: South African National Biodiversity Institute.

Soderberg, K., & Compton, J. S. (2007). Dust as a nutrient source for fynbos ecosystems, South Africa. Ecosystems, 10, 550–561. https://doi.org/10.1007/s10021-007-9032-0

Stock, W. D., & Verboom, G. A. (2012). Phylogenetic ecology of follar N and P concentrations and N:P ratios across mediterranean-type ecosystems. Global Ecology and Biogeography, 21(12), 1147-1156.

Swets, D. L., Reed, B. C., Rowland, J. D., Marko, & S. E., (1999). A weighted least-squares approach to temporal NDVI smoothing. In Proceedings of 1999 Annual Conference: From Image to Information. Bethesda, MA: American Society for Photogrammetry and Remote Sensing.

Verboom, G. A., Moore, T. E., Hoffmann, V., & Cramer, M. D. (2012). The roles of climate and soil nutrients in shaping the life histories of grasses native to the Cape Floristic Region. Plant and Soil, 355, 323–340. https://doi.org/10.1007/s11104-011-1102-6

Vitousek, P. M., Porder, S., Houlton, B. Z., & Chadwick, O. A. (2010). Terrestrial phosphorus limitation: Mechanisms, implications, and nitrogen-phosphorus interactions. Ecological Applications: A Publication of the Ecological Society of America, 20, 5–15. https://doi.org/10.1890/08-0127.1

Yates, C. J., Elith, J., Latimer, A. M., Maitre, D. L. E., Midgley, G. U. Y., Schurr, F. M., & West, A. G. (2010). Projecting climate change impacts on species distributions in megadiverse South African Cape and Southwest Australian Floristic Regions: Opportunities and challenges. Austral Ecology, 35, 374–391. https://doi.org/10.1111/j.1442-9993.2009.02044.x

Zomer, R. J., Trabucco, A., Bossio, D. A., van Straaten, O., & Verchot, L. V. (2008). Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agriculture, Ecosystems and Environment, 126, 67–80. https://doi.org/10.1016/j.agee.2008.01.014

BIOSKETCHES

Cramer is an ecophysiologist interested in the ecophysiological specialization in the hyper-diverse Cape flora and the link between nutrient-impoverished soils and species richness. Verboom works on the assembly of the hyper-diverse Cape flora, its vegetation organization and the role of speciation and extinction in its radiation. However, prior work stresses the ecophysiological and evolutionary basis of the richness of Cape flora, the present work addresses the constraint imposed by the lack of edaphic information and the importance of this for modelling species distributions and understanding species richness, particularly of the Cape flora. Wootton and van Mazijk are both post-graduate students working on species turnover in the Greater Cape Floristic Region.

Author contributions: M.D.C and G.A.V. conceived the study and MDC analysed data and wrote manuscript that was edited by all authors. The data were collated by L.M.W and R.v.M.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Cramer MD, Wootton LM, van Mazijk R, Verboom GA. New regionally modelled soil layers improve prediction of vegetation type relative to that based on global soil models. Divers Distrib. 2019;25:1736–1750. https://doi.org/10.1111/ddi.12973