Vol. 41: 205–220, 2010
doi: 10.3354/cr00846

CLIMATE RESEARCH
Clim Res

Published May 4

Simple statistical approach for computing land cover types and potential natural vegetation

M. Zampieri[1], P. Lionello[2,*]

[1]ISAC-CNR, Lecce, Italy
[2]Department of Material Sciences, University of Salento and CMCC, Italy

ABSTRACT: The main objective of this study was to show the potential of a simple and computationally inexpensive statistical method for the computation of land cover types (LCTs) and potential natural vegetation (PNV), which can be easily adapted to any LCT scheme used by climate models. We propose a diagnostic model (Vegetation Reconstruction by Diagnostic Equilibrium, VERDE), which is based on the cluster analysis of high-resolution datasets of observed LCT distribution and of climate variables. We discuss the reliability of this statistical approach and show that VERDE can be applied for reconstructing PNV distribution in areas such as Europe, India, and China where original vegetation has been replaced by crops and urban areas. According to VERDE, the dominant PNV consists of broadleaf deciduous trees in Central Europe, mixed savanna and grassland in Eastern Europe at mid-latitudes, and evergreen needle trees in Russia. Large areas of India are covered by savanna, and of China by grassland, mixed forest, and evergreen broadleaf trees. VERDE was applied to 5 climate model scenarios (produced by HadCM3, GFDL-CM2.0, IPSL-CM4, CSIRO-MK3, and CNRM-CM3) to identify changes in potential vegetation at a global scale that would be induced by the projected climate change at the end of the 21st century. In the Northern Hemisphere, our results showed an increase in barren soils (deserts) in the areas from the tropics to the mid-latitudes, a northward shift of various types of forest, and a reduction in snow- or ice-covered land and in areas occupied by shrubs and bushes (tundra) at high latitudes. Changes were smaller in the Southern Hemisphere and suggest increases in savanna in South America and shrublands in Australia.

KEY WORDS: Land cover · Potential vegetation · Landuse change · Climate change · K-means clustering · Impacts

1. INTRODUCTION

Land cover types (LCTs) are used for characterizing and modeling the land surface. Many climate models refer to LCTs for computing the interaction between the atmosphere and the land surface, for variables such as surface albedo, surface roughness, momentum and heat fluxes, and evapotranspiration. Vegetation types are connected to LCTs, as vegetation is a key element characterizing the land surface, together with urban areas, lakes, glaciers, and ice caps. The concept of potential natural vegetation (PNV, Tüxen 1956) describes the vegetation types that would exist because of climate, in the absence of a direct human influence such as the introduction of crops and towns, which replace the natural environment that would have otherwise existed. In this study, PNV is a key concept, as our diagnostic model (called Vegetation Reconstruction by Diagnostic Equilibrium, VERDE) addresses the problem of the computation of LCTs in equilibrium with climate and, therefore, also that of PNV.

PNV has been recently used in climate and environmental research to achieve a wide variety of goals: climate characterization, climate change analysis, nature conservation, landscape planning, and ecological risk assessment (e.g. Alo & Wang 2008, Lapola et al. 2008, Rosati et al. 2008). Examples of PNV reconstructions in the literature are based on equilibrium vegetation models (Haxeltine & Prentice 1996) or on a combination of models and observational data (Ramankutty &

*Corresponding author. Email: piero.lionello@unisalento.it

Foley 1999, hereafter referred to as 'RF'). Although there is a consensus on the importance of the concept of PNV, properly defining and identifying it is not without problems (Härdtle 1995), because of problematic issues such as including the effect of permanent site changes (which may be the consequence of human actions) and considering the balance of PNV with all site conditions.

The main tool for the construction of PNV and estimation of climate-related PNV change is the running of a state of the art vegetation model until an equilibrium condition is reached. Dynamic global vegetation models (DGVMs) simulate ecological and physiological processes and are capable of simulating responses of vegetation to climate evolution, also accounting for soil properties and atmospheric CO_2. The models generate predictions of the composition and structure of vegetation for a given climate in terms of relatively few plant functional types, which share similar basic properties regarding their effects on energy, water, and CO_2 exchanges with the atmosphere (Woodward et al. 1995, Haxeltine & Prentice 1996, Sitch et al. 2003). These models are currently used for analyzing climate change impacts on vegetation (Cramer et al. 2001, Alo & Wang 2008). In DGVMs, (1) the definition of the functional types is part of the structure of the model, and (2) the computation of PNV requires a simulation with constant climate conditions and a duration sufficient to reach a steady state of the vegetation.

The main goal of this study was to develop a simple, inexpensive approach for the computation of LCTs on the basis of the association of vegetation types with climate at equilibrium. Here we describe a flexible (to be used with any reasonable vegetation or land-cover classification dataset), user-friendly (easy to implement), and inexpensive (no large computer resources are required) model, called VERDE. With respect to DGVMs, VERDE avoids the rigidity of point (1) above and the need of (2).

VERDE is an application that involves cluster analysis. Cluster algorithms aim to group objects by an automatic and objective procedure. The idea is to identify, within a sparse distribution, objects that are close to each other and can be represented by a centroid associated with each cluster. The algorithm requires a definition of distance for describing differences between elements. In this study, the technique was applied to LCTs and to the annual cycle of monthly precipitation and temperature, but it could be adapted to include other meaningful climate variables such as solar radiation, daily temperature range, growing degree days, moisture indices, and temperatures of the warmest/coldest months, as well as non-climatic fields such as soil types. In the literature, some examples of cluster techniques have been applied to the classification of regional climate (e.g. Unal et al. 2003) and to confirm or replace the subjective Köppen climate classification (Köppen 1900), which has been widely used in the past, mainly by geographers (Kottek et al. 2006, see Peel et al. 2007 for a recent analysis). A recent study (Wang & Price 2007) implemented the same technique to annual climatic indices and to a small subset of plant functional types for classification purposes. Moreover, to our knowledge, the cluster technique has never been used for modeling LCTs and their climate change-related changes.

The approach is data driven, as VERDE is based on information that must be provided by existing datasets containing LCTs and climate variables. The existence of an underlying dynamic that associates LCTs with climate variables is conceptually needed to support the statistical links that VERDE exploits, but no dynamical relation is included in the model and no explicit knowledge of the prognostic relation between LCTs and climate is needed for building VERDE. In synthesis, VERDE is a diagnostic tool for associating LCTs to climate variables on the basis of cluster analysis of large datasets. VERDE is usable only for 'natural' LCTs; urban areas, crops, and farmland cannot be analyzed with this simple approach.

VERDE uses the observed links between LCTs and climate for computing the spatial distribution of LCTs. It aims at describing only the final equilibrium land cover condition, without any attempt at describing its time evolution. VERDE cannot be an alternative to DGVMs, not only because it does not compute transient conditions, but also because it does not provide information on variables such as net primary production, carbon content in vegetation and soil, carbon fluxes, or evapotranspiration. In the implementation described in this study, VERDE is targeted at the representation of LCTs that are used in climate models for the purpose of describing the surface properties that are important for computing fluxes and atmospheric circulation.

Section 2 contains a technical description of VERDE, of the climate and land-cover data used in its cluster algorithm, and a model validation through the reconstruction of LCTs over the Americas. Section 3 describes the PNV that VERDE reconstructs over strongly anthropized regions and a set of 5 global LCT distributions, which result from the temperature and precipitation patterns produced by the global model simulations of HadCM3, GFDL-CM2.0, IPSL-CM4, CSIRO-MK3, and CNRM-CM3 for the A2 scenario (Nakićenović & Swart 2000). Section 4 contains an overall discussion of the results, the limitations of VERDE, its advantages, and how it compares to DGVMs and other diagnostic descriptions of land cover. Section 5 is a short synthesis of the main conclusions of this study. Mathematical details related to the definition of the clustering procedure are given in Appendix 1, where specific technical features adopted for the clustering are also described. Appendix 2 contains some technical details on the procedure used for comparing the results of VERDE to the PNV proposed by RF.

2. DATA, METHODOLOGY, AND MODEL VALIDATION

2.1. Vegetation and climate data input

As the idea of the model is to use global data to build a purely statistical association of LCT in equilibrium with climate, 2 reliable sets of input data are needed: a global LCT, or vegetation type distribution, and a global climate dataset.

Although the approach can be adapted to different choices of LCTs, all applications in this study use the Global Land Cover Characteristics (GLCC) International Geosphere Biosphere Programme (IGBP) data set, which contains a global land use fractional distribution at 10′ resolution derived from 1 km AVHRR observations for the period April 1992 to March 1993 (Loveland et al. 2000) with 17 classification types. We ignored 5 of these (wetland, cropland, urban, water, cropland-natural), because they are not directly associated with climate. The remaining number of vegetation types (N_V) = 12 LCTs were used for VERDE: barren soil, evergreen broadleaf trees, deciduous broadleaf trees, evergreen needle trees, deciduous needle trees, grassland, savanna, woody savanna, mixed forest, open shrubs, closed shrubs, and snow-ice. The land cover at each grid point is distributed percentage-wise among the various types. Although the LCT with the largest percentage, which is called the dominant LCT, is often used to visualize and discuss results, the simultaneous presence of several LCTs is fully accounted for and exploited by VERDE.

When building the statistical association, it is important to consider only natural LCTs whose distribution is not altered by human influence and factors other than climate. Therefore, only points with ≤5% of croplands, ≤5% urban and built-up areas, ≤5% water bodies, ≤5% permanent wetlands, and ≤5% cropland-natural vegetation mosaics were considered, so that the points used by VERDE have a minimum of 75% natural LCTs.

In the applications shown in this study, a minimal set of climate variables was used, consisting of the annual cycle of monthly mean temperature and accumulated precipitation derived from the data provided by the Climatic Research Unit (CRU) of the University of East Anglia (New et al. 2002, www.cru.uea.ac.uk/cru/data/) at 10′ resolution. Further refinement considering more climate variables and also non-climatic fields will be considered in future applications.

2.2. VERDE model approach

In VERDE, cluster definition is based on the average monthly temperature (t) and precipitation (p) annual cycle in each point i,j of the grid: $t^{(m)}$, $p^{(m)}$, m = 1, 12 — where m represents the calendar month — and the percentage of land cover for vegetation types, $v^{(n)}$, m = 1, N_V — where n represents the LCT (the grid row and column indices i,j are omitted for brevity of notation). The equivalent dimensionless variables are defined as

$$\tilde{t}^{(m)} = \frac{t^{(m)}}{\mathrm{SD}_T} \quad m = 1,12 \tag{1}$$

$$\tilde{p}^{(m)} = \frac{p^{(m)}}{\mathrm{SD}_P} \quad m = 1,12 \tag{2}$$

$$\tilde{v}^{(n)} = \frac{v^{(n)}}{\mathrm{SD}_V} \quad m = 1, N_V \tag{3}$$

where SD_T, SD_P, and SD_V are the corresponding standard deviations computed on all grid points and either months or vegetation types. In this manuscript, the tilde denotes dimensionless variables (see Appendix 1), lowercase characters denote the values at a single grid point, and uppercase characters denote average or global values. The dimensionless variables are used for the definition of $\tilde{x} = (\tilde{t}, \tilde{p}, \tilde{v})$, which is a dimensionless array of size N_D = 12 + 12 + N_V. By means of a k-means cluster algorithm (MacQueen 1967; see Appendix 1 for more details), $\tilde{x}_{il}$ and $\tilde{x}_{C_l}$ are identified and represent the 'position' of the ith element of cluster l and the centroid (baricentrum) of cluster l, respectively. The advantage of the dimensionless variables is that the clustering algorithm gives identical weight to precipitation, temperature, and vegetation types in the computation of clusters.

VERDE adopts a 2-step method for the attribution of LCT on the basis of the annual cycle of precipitation and temperature. Step 1 (analysis, based on observations) is the cluster analysis of the LCTs and climate data and results in the definition of the model clusters. Step 2 (attribution, based on processing of observed or simulated climate data) assigns vegetation types to each grid point. During the attribution step, the local annual cycle is computed for each point of the domain, and the cluster with the most similar annual cycle is identified by finding the nearest centroid in the subspace scanned by the variables $\tilde{t}^{(m)}$ and $\tilde{p}^{(m)}$ for m = 1, 12. The distance from the centroid in such subspace is a measure of how well the cluster represents the local climate. The LCT vector $(\tilde{v}^{(1)}, \tilde{v}^{(2)} \ldots \tilde{v}^{(N_V)})$ of the cluster centroid is attributed to the point.

Wang & Price (2007) already tested the k-means cluster analysis and found an optimal classification tool, but it was applied for characterizing 3 plant functional types and was applied to only 3 climatic indices: monthly mean daily minimum temperature, annual sum of growing degree days where daily mean temperature exceeded 5.0°C, and the climatic moisture index (which was defined as total annual precipitation minus annual potential evapotranspiration). Wetlands were excluded from the analysis by Wang & Price (2007) and also from our study. Here, we cluster a complete LCT distribution in association with climate (given as the annual cycle of monthly temperature and precipitation) and use the results for a diagnosis of vegetation coverage.

2.3. Model tuning and optimization

The resulting distribution of the attributed LCTs depends on the number of clusters that has been prescribed (Fig. 1 shows the dominant LCT). In the simplest configuration (2 clusters), the whole land surface is split between barren soil and open shrubs. With increasing numbers of clusters, the dominant surface cover becomes progressively diversified. With 3 clusters, evergreen broadleaf trees replace barren soil in the humid tropical regions. With 4 clusters, woody savanna appears over a large fraction of transitional areas around the desert regions. With 5 clusters, open shrubs replace many barren soil areas and are added in transitional regions between savanna and barren soil. With 6 clusters, snow-ice covered regions are identified at high latitudes in the Northern Hemisphere. With 7 clusters, mixed forests occupy large parts of Eurasia and North America. With 8 clusters, grassland covers large regions at mid-latitudes in the northern Hemisphere. With 9 clusters, a distinction between areas with woody savanna and savanna is introduced. Increasing the number of clusters does not necessarily introduce a new dominant LCT (there are only 12 LCTs in this implementation), but is meant to provide a finer classification. In fact, no new dominant LCT appears with 10 clusters; rather, only a geographical redistribution of dominant LCTs takes place (e.g. barren soils replacing open shrubs at high latitudes) with respect to that based on 9 clusters. Moreover, the LCTs in the GLCC dataset are not univocally linked with climate, as quite different annual cycles can be associated to the same LCTs. An example is barren soil, which can be linked to the desert tropical dry climate and to the tundra landscape of the polar regions.

In the k-means cluster algorithm, the number of clusters is prescribed and can be changed freely by the user. Increasing the number of clusters implies improving VERDE's capability of describing the actual LCT distribution as a function of climate. At the same time, the computation of the clusters becomes progressively more time consuming, and the conceptual classification and simplification of the link between climate and vegetation becomes less effective. Appendix 1 discusses the optimal number of clusters in the range up to 500 clusters and reaches the conclusion that it would not be advisable to use fewer than about 100 clusters.

Experimenting with the VERDE model has shown that the procedure can produce unreliable results in connection with the presence of clusters with few elements, which are clusters attributed to few grid points (small areas). When the analysis (the first step of the model procedure) contains small clusters, the attribution (the second step of the procedure) on the basis of the local annual temperature and precipitation cycle can assign their LCTs to large areas, and such attribution depends irregularly on the number of clusters. This problem has been avoided by excluding clusters smaller than 0.4 times the average cluster size during the attribution (see Appendix 1). Moreover, note that the meaning of small clusters is arguable, as they might represent the effect of local factors or climatically non-equilibrium situations, such as remnants of LCTs that are disappearing.

For all figures and results presented in this manuscript, VERDE was implemented using 100 clusters, and the attribution did not consider clusters smaller than 0.4 times the average cluster size. Fig. 2c shows the natural LCT distribution computed by VERDE in this configuration.

2.4. Model validation

Fig. 2 shows the capability of VERDE to describe the actual dominant LCT. Fig. 2a shows the LCT in the global GLCC IGBP dataset. Fig. 2b shows only the points where the LCT has been considered natural (the anthropized grid points are masked). Fig. 2c shows the dominant LCT of the centroid of the cluster to which each single grid point is attributed. The natural LCT in Fig. 2b matches well the LCT of the cluster centroid in Fig. 2c, showing the consistency of the model and the capability of the cluster centroids to correctly represent the observed natural LCT and its link with climate. The kappa score (Cohen 1960) value for the agreement between Fig. 2b and c is 0.62, suggesting a good agreement between the 2 datasets.

In order to validate VERDE, the natural vegetation over the Americas is reconstructed using the clusters computed from the data (temperature, precipitation, and vegetation) over the rest of the globe, but excluding those over the American continent itself. Six maps are shown in Fig. 3. Panels (a), (b), and (c) show the actual LCT, the actual natural LCT, and the reconstructed LCT, respectively. Panel (d) shows the distance: in the subspace scanned by $\tilde{t}_{ij}^{(m)}$ and $\tilde{p}_{ij}^{(m)}$ between each point *ij* and the centroid C_{ij} of the cluster

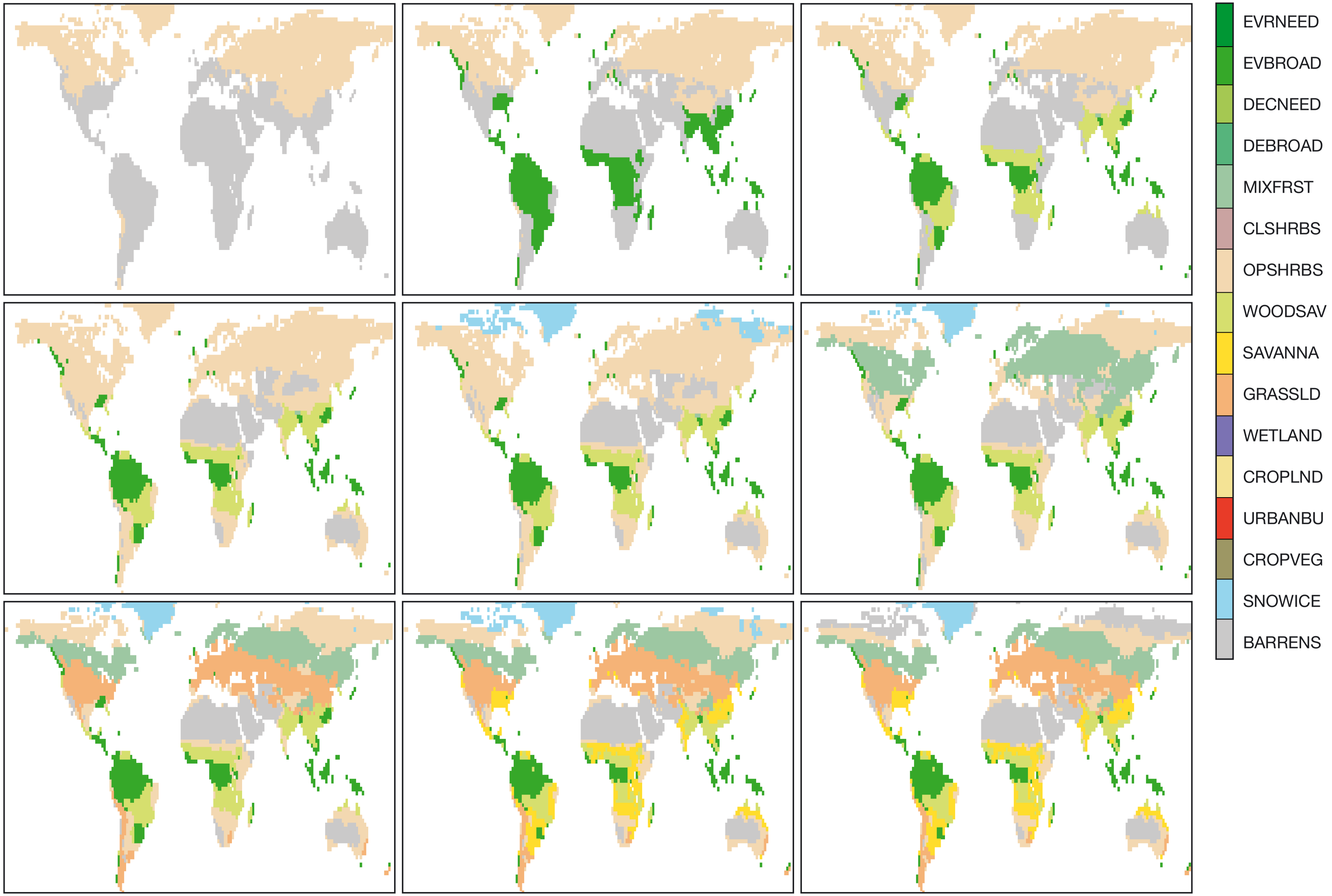

Fig. 1. Dominant vegetation types with increasing number of clusters: from left to right and from top to bottom, 2, 3, 4, 5, 6, 7, 8, 9, 10 clusters are used. For plotting purposes, distribution values of reconstructed potential natural vegetation (PNV) are averaged on a 2° grid. Land cover types—EVRNEED: evergreen needle trees, EVBROAD: evergreen broadleaf trees, DECNEED: deciduous needle trees, DEBROAD: deciduous broadleaf tress, MIXFRST: mixed forest, CLSHRBS: closed shrubs, OPSHRBS: open shrubs, WOODSAV: woody savanna, SAVANNA: savanna, GRASSLD: grasslands, SNOWICE: snow ice, BARRENS: barren soil

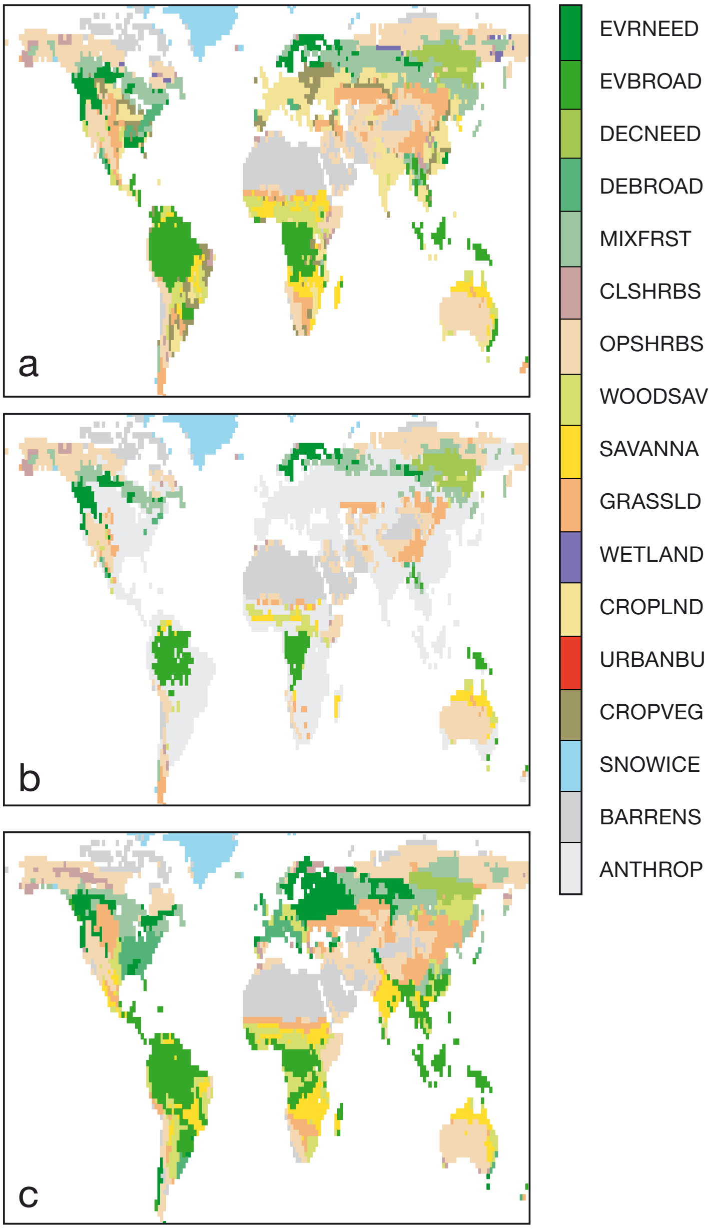

Fig. 2. (a) Actual dominant vegetation type based on the GLCC-IGBP dataset (Loveland et al. 2000); (b) same as (a) except that only the points with natural vegetation are plotted and anthropized regions are shaded in light gray; (c) dominant potential natural vegetation (PNV) of cluster centroids, using 100 clusters. PNV values are aggregated on a 2° grid

containing the grid point itself. Low/high values of $\text{dist}_{TP_{ij}}$ are a measure of the success/failure of the attribution procedure, that is of the identification of a climate outside the Americas sufficiently similar to the local one. Fig. 3e provides information on the error resulting from matching the annual temperature cycle of the cluster centroid to the grid point value. This is a dimensional quantity (values in K) given by the deviation around the temperature of the cluster centroid. Fig. 3f shows the analogous quantity for precipitation (in mm). Formulas describing these quantities are given at the end of Appendix 1.

Discrepancies between actual natural and reconstructed LCTs (shown in Fig. 3b,c) are related mostly to an overestimate of the area attributed to grassland (which replaces open shrubs) and savanna and woody-savanna, which replace grassland and open shrubs over central areas of North America. Comparable errors take place in the southernmost part of South America and along the Andes. The evergreen needle forest along the north Pacific coast is partially replaced by evergreen broadleaf trees. The map of the distance (Fig. 3d) shows that all of these inaccuracies are due to failures in identifying a cluster centroid with a similar climate. This is likely due to the absence of similar climatic conditions in other regions of the world. Peculiarities of the annual cycle of temperature (Fig. 3e shows areas where errors in temperature are rather large) are the likely candidates for these discrepancies. A different error is the introduction at high latitudes of large patches of deciduous needle trees that are not observed in the present vegetation. This is due to the presence of large extensions of deciduous needle trees over continental cold regions of Asia with similar climate conditions. Differences between the reconstruction and observations are also large in southern Greenland, where tundra replaces the existing glaciers. In fact, another possible cause of wrong attribution is the lack of dynamics in VERDE, which in principle computes LCTs in equilibrium with climate. Obviously, the presently observed LCTs could represent a metastable or transitional state, which is not in equilibrium and it is gradually disappearing or owing its existence to a microclimatic balance. Any metastable or transient condition would be completely missed by VERDE. This could be the source of the differences between attributed and present LCTs in southern Greenland. In spite of these shortcomings, over most of the Americas, the attribution is successful because the present LCTs are confirmed by the attribution, and crops are replaced with plausible LCTs. In fact, the kappa score computed between the natural and reconstructed LCT distributions is 0.48. Therefore, Fig. 3 shows that information on the annual temperature and precipitation cycle alone is sufficient for identifying the actual LCT over the Americas on the basis of statistical links established using data from outside the Americas.

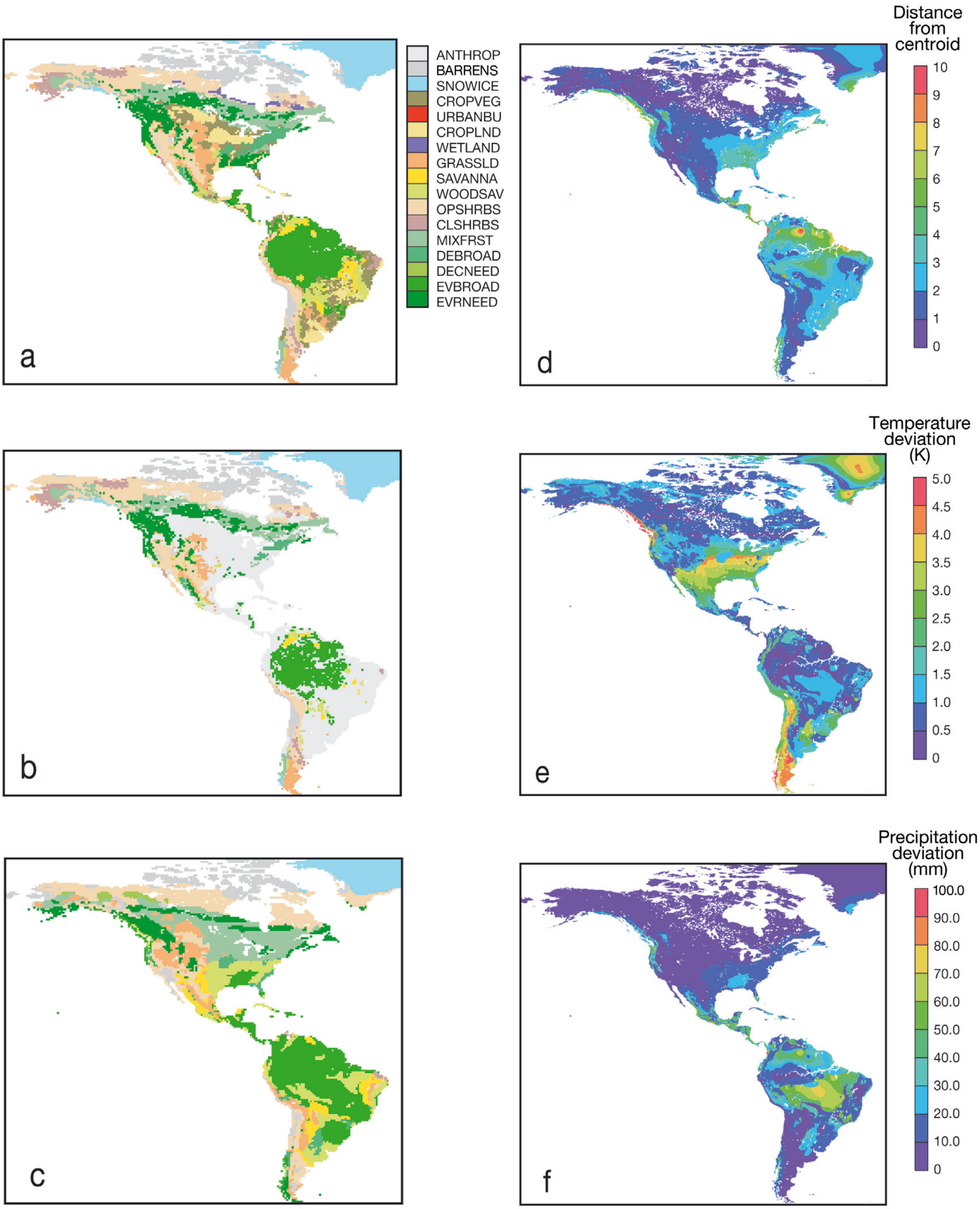

Fig. 3. (a) Observed vegetation, (b) observed natural vegetation, (c) model potential natural vegetation (PNV), (d) distance from cluster centroid, (e) and (f) temperature (K) and precipitation (mm) deviation, respectively, from that of the cluster centroid. PNV values are aggregated on a 40′ grid

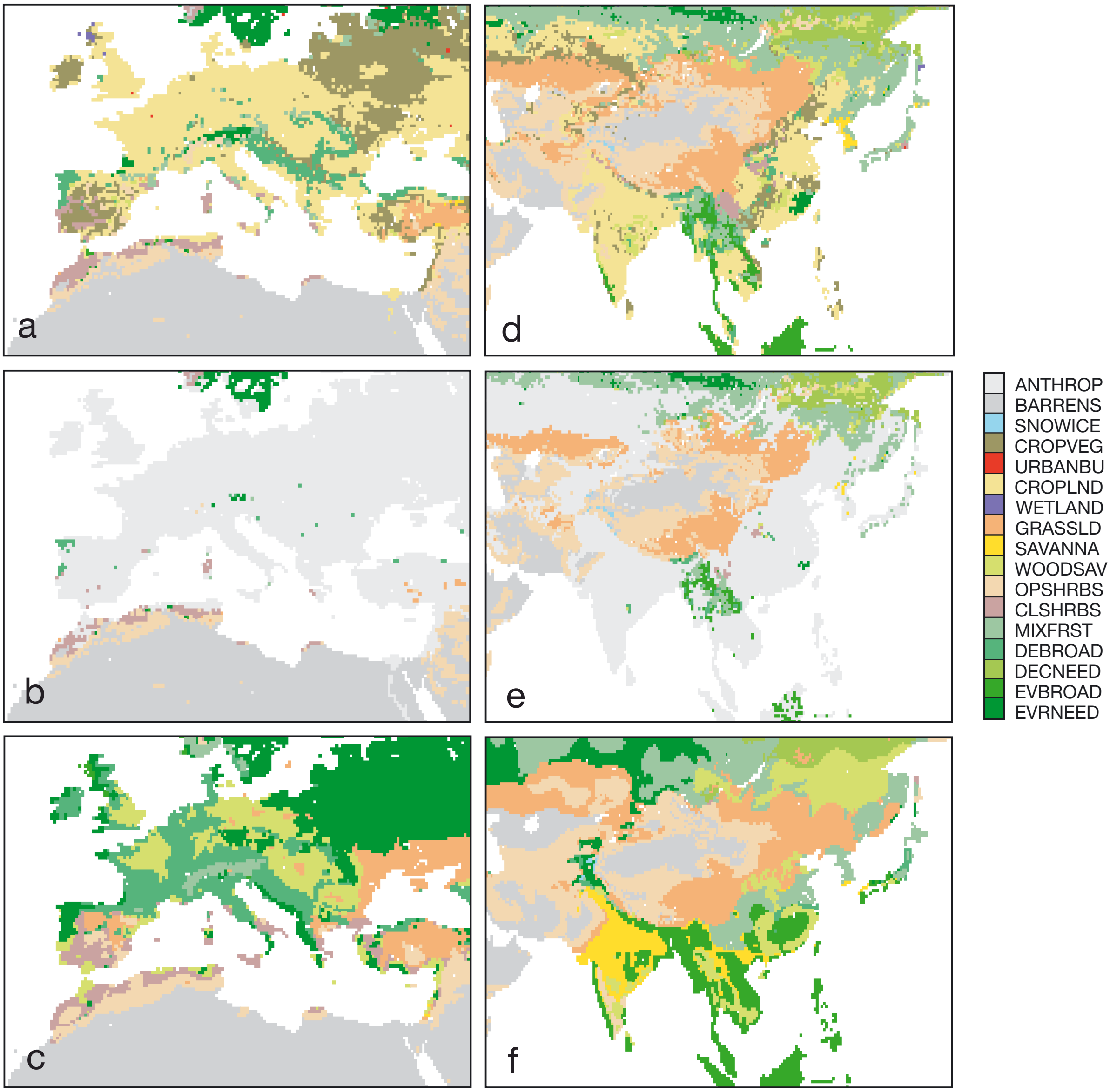

Fig. 4. (a,d) Observed vegetation, (b,e) observed natural vegetation, and (c,f) potential natural vegetation (PNV) over Europe (a–c) and over Asia (d–f). The local annual cycles of temperature and precipitation are used in each point for attributing to it the potential natural vegetation. Distribution values are aggregated on a 20′ grid for Europe and on a 30′ grid for Asia

3. APPLICATIONS AND RESULTS

This section describes 2 applications of VERDE that demonstrate its effectiveness: the reconstruction of PNV and the computation of LCTs in climate projections.

3.1. Reconstruction of PNV over anthropized areas

An application of VERDE is the reconstruction of PNV types over areas where farming, urban areas, and other human-related activities have replaced the original vegetation cover. The best examples are the almost completely anthropized areas in Europe, China, and India. Over these areas, the LCT reconstruction obtained using VERDE can be considered an independent contribution to the reconstructions produced by botanists (e.g. Neuhausl 1991) and to the general discussion on PNV (Moravec 1998, Härdtle 1995). In general, botanists follow a completely different reasoning and have different targets, so that a precise comparison is not possible. However, VERDE qualitatively reproduces the reconstruction by Neuhausl (1991). Fig. 4 compares the actual present vegetation and natural vegetation over Europe and Asia to the PNV over Europe and Asia. The dominant PNV over most of central Europe consists of deciduous broadleaf trees, with large patches of mixed wood-savanna vegetation appearing towards the north-east. Evergreen needle trees dominate over Russia and Scandinavia. Shrubs and barren soil (desert) are located along the southern coasts of the Mediterranean. Grasslands occupy the interior of Anatolia, part of Iberia, and the northern coast of the Black Sea.

In Asia, all of India, China, and large parts of western Asia have no natural vegetation left. According to VERDE, India would be mostly savanna, with evergreen broadleaf trees occupying only its humid western coast and most of Indonesia. China would be covered with grassland, mixed forest, and evergreen broadleaf trees, with transition among these types occurring from north to south. Western Asia, located at higher latitudes than China, would be covered with mixed forest, evergreen needle trees, and grassland, with a patchy distribution roughly corresponding to a transition from north to south.

Reconstructed vegetation is consistent with the LPJ (Lund Potsdam Jena) model results (Sitch et al. 2003) at the global scale, although a precise comparison is not possible, because in these applications, VERDE uses an LCT classification different from that of LPJ. Compared to LPJ, VERDE underestimates the amount of evergreen and deciduous needle trees in Central Europe (France, Germany, and England) and also over Iberian and Italian orographic features.

The validity of VERDE has been checked against the RF reconstruction that is available at www.sage.wisc.edu/atlas/ (Fig. 5a). Since RF did not adopt the GLCC IGBP LCT classification, some extra rules had to be introduced in order to transform the VERDE classification to the RF vegetation types (see Appendix 2). The transformed VERDE classification is shown in Fig. 5b. The 2 maps (Fig. 5a,b) match reasonably well with a kappa score of 0.41, denoting ‘moderate agreement.’ A difference between them is the larger tundra extension in VERDE in areas where RF set mixed and boreal forest. The source of this VERDE attribution is the presence in the GLCC IGPP map (Fig. 2a) of natural open shrubs in cold regions, which are considered tundra with the conversion rules in Appendix 2. The presence of savanna in Europe in the VERDE maps is not correct, but is explicable, as RF also set savanna over 2 tiny spots in Europe and at similar latitudes over central Asia. Finally, VERDE attributes savanna as natural vegetation over India because the Indian climate is assigned to the same cluster as areas of Australia, Africa, and South America that are located at the same distance from the Equator, but in the southern hemisphere, and where GLCC IGBP maps show savanna. Thus, the presence of differences between VERDE and RF is not unusual in this sort of comparison. A recent example can be found in Lapola et al. (2008), who compared RF to the PNV by Matthews (1983). They were also forced to establish conversion rules between the 2 sets of vegetation types and found a kappa score of 0.49.

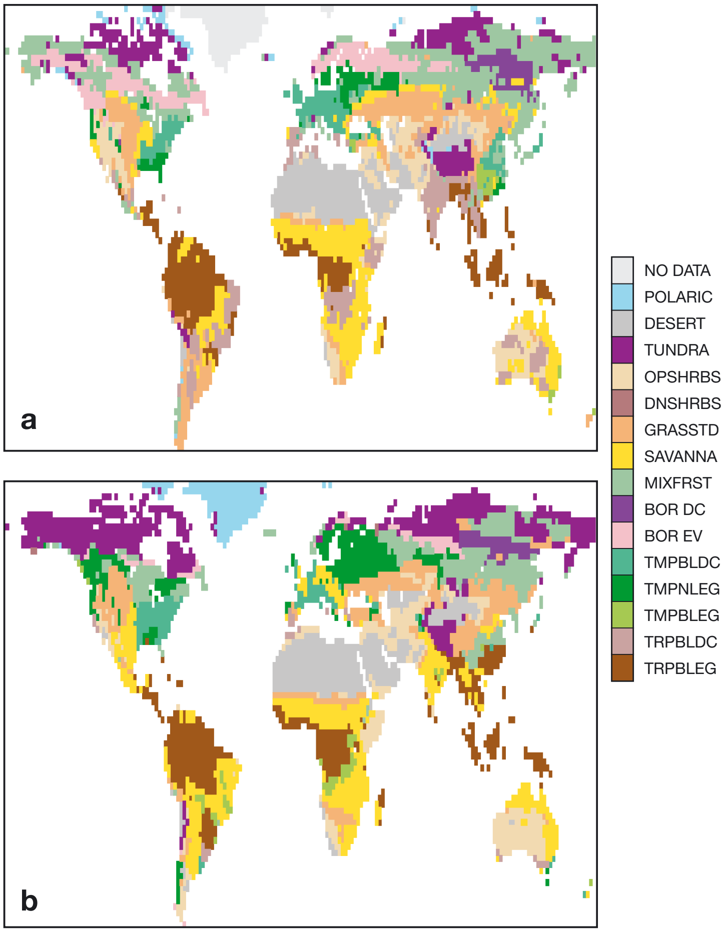

Fig. 5. Global maps showing the potential natural vegetation (PNV) at 2° resolution. (a) Results of Ramankutty & Foley (1999); (b) results produced by our model, VERDE

3.2. Effects of climate change on LCTs

Another application of VERDE is the computation of climate-induced LCT change. This was carried out using the results of global climate simulations and considering the mean annual cycle for the CTR (ConTRol) period 1961–1990 (CTR) and for the period 2071–2100 of the A2

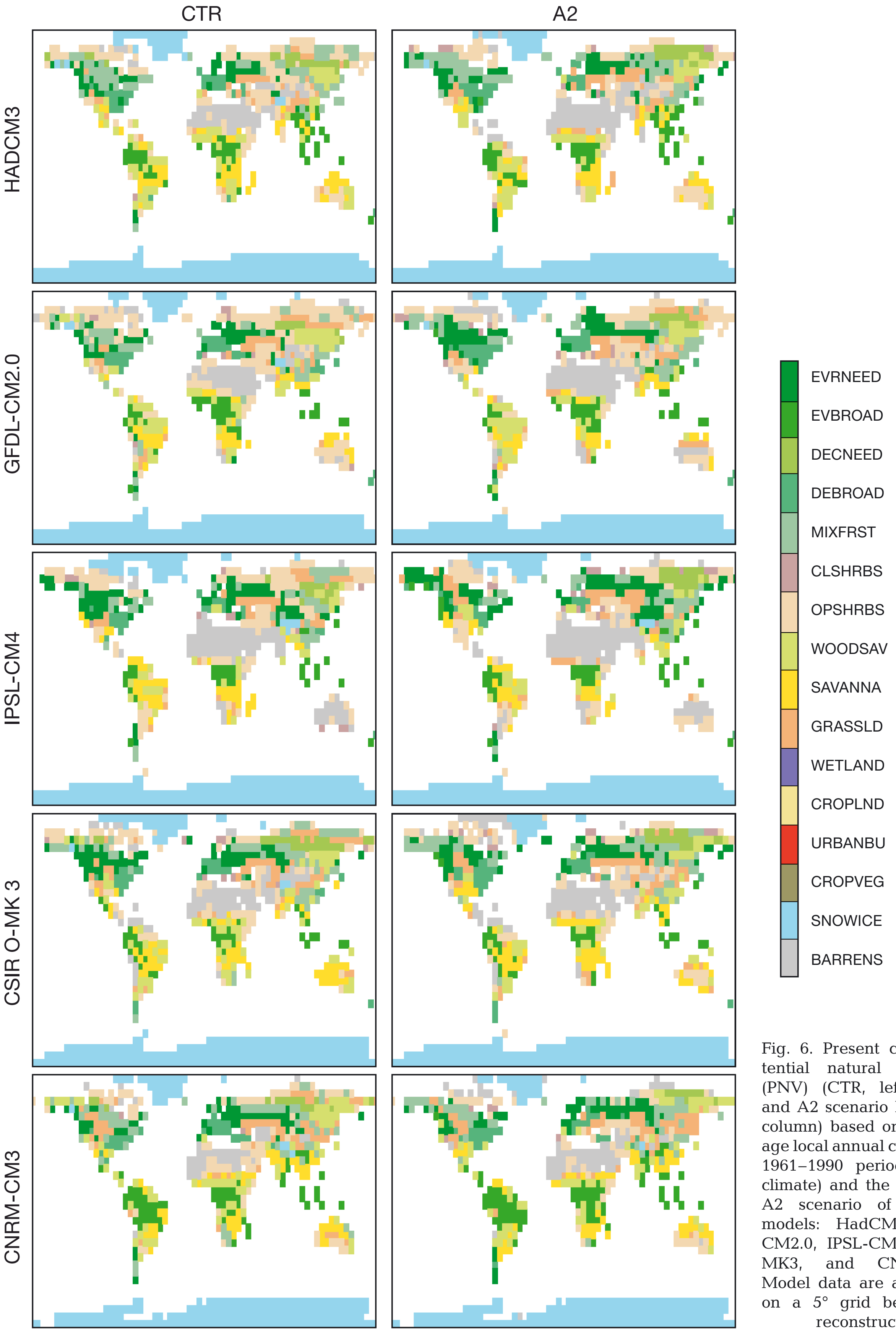

Fig. 6. Present climate potential natural vegetation (PNV) (CTR, left column) and A2 scenario PNV (right column) based on the average local annual cycle for the 1961–1990 period (present climate) and the 2071–2100 A2 scenario of 5 global models: HadCM3, GFDL-CM2.0, IPSL-CM4, CSIRO-MK3, and CNRM-CM3. Model data are aggregated on a 5° grid before PNV reconstruction

emission scenario. Fig. 6 shows the LCT change derived from 5 model simulations (HadCM3, GFDL-CM2.0, IPSL-CM4, CSIRO-MK3, CNRM-CM3), whose results were extracted from the CERA WWW-Gateway of the World Data Center for Climate, Hamburg (http://cera-www.dkrz.de/). Model resolution varies from about 1.9° (the spectral T63 CNRM-CM3 and CSIRO-MK3 models) to 3.75° (the longitude resolution of HadCM3 and IPSL-CM4). In this application of VERDE, all models were transferred to a common 5° resolution latitude-longitude grid.

A feature common to all of these climate projections is the northward shift of the forest areas in the Northern Hemisphere at the expense of barren or snow-ice covered soil and shrubs. A second feature is the increased extension, both north- and southwards, of the barren tropical areas of the Northern Hemisphere. Those changes are associated with milder temperature conditions at high latitudes and with a northward and southward extension of areas with very scarce precipitation in the tropics, respectively. Changes in the Southern Hemisphere are smaller and less consistent among different models.

Table 1 summarizes the changes of LCTs, showing the percentage of points for each LCT that VERDE attributes during the CTR period, differences relative to the A2 scenario (A2-CTR), and the corresponding change percentage, i.e. the difference divided by the value in the CTR period. Values of reconstructed LCTs for the CTR period are consistent among the models, and the k-scores compared to the natural LCTs of VERDE range from 0.40 to 0.55. According to this diagnostic, the best model among those we took into account is HadCM3; all others had k-scores below 0.44. Some differences produced by the A2 scenario are important, including the increment of barren soil (+26%) and the decrease in open shrubs (–15%) and snow-ice covered land (–4.5%). The areas covered by other LCTs are shifted in the A2 scenario (mainly in the Northern Hemisphere and northward) but their overall extension does not really change. This suggests that the level of the related resources will remain steady in the future climate, although the areas where they are located will change. In contrast, the increase in barren soils represents an overall loss of productivity.

Table 1. CTR (ConTRol): present climate potential natural vegetation (PNV) mean distribution (%) computed over all land points based on the average local annual cycle for 1961–1990 (present climate) of 5 global models (HadCM3, GFDL-CM2.0, IPSL-CM4, CSIRO-MK3, and CNRM-CM3). A2 – CTR: difference between the 2071–2100 A2 scenario and the 1961 to 1990 period; (A2 – CTR)/CTR: as for 'A2 – CTR', but for the relative difference (percentage change of PNV). Land cover types: evergreen needle trees (EN), evergreen broadleaf trees (EB), deciduous needle trees (DN), deciduous broadleaf tress (DB). mixed forest (MF), closed shrubs (CS), open shrubs (OS), woody savanna (WS), savanna (S), grasslands (G), snow-ice (SI), barren soil (BS)

	EN	EB	DN	DB	MF	CS	OS	WS	S	G	SI	BS
CTR												
HadCM3	5.3	6.3	1.7	3.7	9.5	0.7	15.2	6.9	6.3	5.0	31.4	8.1
GFDL-CM2.0	5.7	4.5	1.0	4.2	7.2	1.6	17.7	6.6	5.4	5.6	31.3	9.4
IPSL-CM4	9.0	3.9	1.4	3.0	6.7	1.8	14.9	3.8	5.1	4.7	32.2	13.4
CSIRO-MK3	7.4	4.4	2.6	3.7	7.6	1.5	12.6	6.7	6.2	6.3	31.5	9.5
CNRM-CM3	6.5	8.2	1.8	4.1	8.7	0.9	11.4	7.7	6.6	8.0	29.9	6.2
Mean	6.8	5.5	1.7	3.8	7.9	1.3	14.3	6.3	5.9	5.9	31.2	9.3
SD	1.3	1.6	0.5	0.4	1.0	0.4	2.2	1.3	0.6	1.2	0.7	2.4
A2 – CTR												
HadCM3	1.5	0.2	0.3	0.1	–0.3	0.9	–2.9	–0.4	–0.6	0.4	–1.1	2.0
GFDL-CM2.0	1.8	0.1	0.4	0.5	–0.1	0.2	–4.1	–0.4	–0.6	0.3	–1.0	2.7
IPSL-CM4	1.7	1.0	0.8	–0.2	–0.7	0.4	–2.8	0.4	–0.9	1.1	–1.6	0.9
CSIRO-MK3	–0.1	0.1	–0.8	0.9	0.5	0.4	–0.8	–0.9	0.4	0.5	–2.3	2.2
CNRM-CM3	–0.1	1.1	–0.8	0.7	–0.3	0.2	–1.1	–0.4	0.1	–1.4	–1.0	3.0
Mean	1.0	0.5	0.0	0.4	–0.2	0.4	–2.3	–0.4	–0.3	0.2	–1.4	2.2
SD	0.9	0.5	0.7	0.4	0.4	0.3	1.2	0.4	0.5	0.8	0.5	0.7
(A2–CTR)/CTR												
HadCM3	28.7	2.5	19.9	2.4	–3.5	133.9	–19.1	–6.2	–10.1	7.3	–3.6	24.7
GFDL-CM2.0	32.2	2.8	38.0	11.9	–1.1	13.6	–23.2	–6.1	–10.5	6.1	–3.1	28.8
IPSL-CM4	19.0	26.3	55.5	–7.6	–10.3	23.0	–18.9	9.8	–18.0	23.0	–5.0	6.4
CSIRO-MK3	–0.8	1.2	–28.4	23.0	6.2	26.1	–6.3	–13.6	6.6	7.3	–7.3	23.1
CNRM-CM3	–2.1	13.0	–45.5	17.5	–3.1	20.5	–9.6	–5.7	1.9	–17.1	–3.4	48.9
Mean	15.4	9.1	7.9	9.4	–2.4	43.4	–15.4	–4.3	–6.0	5.3	–4.5	26.4
SD	14.4	9.6	38.7	10.9	5.3	45.4	6.4	7.7	8.9	12.8	1.6	13.6

4. DISCUSSION

The aim of this study was to develop a simple diagnostic approach, called VERDE, for computing LCTs in equilibrium with the local climate by exploiting information existing in large data sets. VERDE adopts a data-driven approach, which is based on statistical links and not on vegetation dynamics. The approach has its utility, because it allows computing the response of LCTs to climate change on long time scales (ideally, VERDE computes the final equilibrium condition) without the need of long forward time integrations. Since many LCTs have a correspondence with vegetation types, VERDE fulfills a similar task for vegetation types and computes PNV.

The VERDE approach has 2 limitations. One concerns the reproduction of present natural LCT distribution; another is its applicability to future conditions.

The underlying assumption of VERDE is that current vegetation types in non-anthropized areas equal PNV and that vegetation is in equilibrium with its environment. Clusters are defined on this basis. However, human influences can be detected almost everywhere, even in apparently 'natural' areas: nitrogen inputs reach considerable magnitude, forestry activities dominate a great proportion of forests, atmospheric CO_2 concentrations are far from pre-industrial levels, and temperatures have increased over the last decades. There is the concrete possibility that natural vegetation in the GLCC IGBP dataset is not in equilibrium with its environment, and in many regions, VERDE's basic assumption might be violated. However, most anthropic factors have only recently reached the scale at which they produce global effects, and systems have begun a transition phase toward new conditions, but they have not yet greatly drifted away from the previous equilibrium.

When projecting LCTs into the future, VERDE assumes constant relationships between the LTCs and climate conditions. However, physiological reactions of plants to altered environmental conditions can be expected on a large scale. Effects of increased atmospheric CO_2 concentrations on plant water-use efficiency can alter the precipitation-vegetation relationship, which can limit VERDE's applicability to future climates. This issue is not unusual in climate change studies. Note that dynamical vegetation models, which explicitly describe the processes responsible for climate evolution, present limitations to their range of predictability as well, because they include parameterizations of plant-physiological processes that have been tuned for the present climate conditions and should be changed when applied in much different future conditions. Similarly, the relations that are at the basis of VERDE and of other statistical models would not be valid if climate conditions drifted too far from those at present. Common scientific sense suggests that the range of predictability of dynamical models is larger than that of statistical models, but both approaches have limitations to their range of applicability.

It is clear that VERDE cannot aim to be an alternative to DGVMs, because it is not able to describe the time-dependent behavior of vegetation communities, to simulate time-dependent biophysical and biogeochemical feedbacks, to describe the terrestrial carbon sinks in future climate scenarios, and to address many of the current research concerns. The advantages of VERDE are its simplicity (both in concept and in implementation) and the freedom it allows in choosing the LCT classification to be used. This is because of its data-driven approach, so that VERDE does not need to be based on LCTs with well known dynamics and prognostic equations. In fact, in this study, the adopted classification followed that adopted by climate models, which are designed to describe surface processes and not vegetation types. Besides the applications described here, which are based on the International Geosphere Biosphere Program (IGBP) classification, VERDE has also been tested with the Biosphere Atmosphere Transfer Scheme (BATS, Yang & Dickinson 1996) and with the classification adopted by the ORCHIDEE model (Krinner et al. 2005) in order to interface VERDE with the regional model Regcm and with the global model LMDZ, which have respectively adopted these 2 schemes (M. Zampieri & P. Lionello unpublished, M. Zampieri unpublished).

Without introducing new methodological features, VERDE can become more accurate by including more climate variables (besides the annual cycle of temperature and precipitation), such as growing degree days, moisture indices, and temperatures of the warmest/coldest months — which are representative of environmental factors to which vegetation is exposed — if those data are available at a global scale. However, the VERDE validation and the applications described in this paper show that adequate results can < obtained using only the annual cycle of temperature and precipitation. In fact, the model results are consistent with the physiological bounds on minimum temperature reported by Haxeltine & Prentice (1996). If, using the CRU dataset, minimum temperature is defined as the mean monthly temperature minus half the diurnal range, then there are no clusters with >10 % of needle trees and a minimum temperature below –60°C or above 0°C and equally no cluster with >10 % of evergreen broadleaf trees and minimum temperature below –10°C.

Note that vegetation is not only limited by the average annual cycle of monthly precipitation and temperature. Extreme climate events can also play a role in determining the vegetation and limiting existing types of plants, both with direct (e.g. drought) or indirect (such as the impact of fires triggered by extremely hot and dry spells) climate effects. The association of droughts and extreme climate events with the annual cycle is definitely weak, so that the effect of these factors is not accounted for by this approach. However, the results of VERDE are consistent with Bond et al. (2004), who included fire effects in a dynamical vegetation model, and one may consider including these effects if such information becomes available, as this is not an intrinsic limitation of the approach adopted by VERDE.

Advantages of VERDE with respect to previously widely used equilibrium models are its objective data-driven approach and its flexibility.

VERDE's approach does not require any *a priori* knowledge of plant physiology and of land cover dynamics, and it is interesting that its results compare favorably to those produced by expert ecologists and climatologists, which are based on other criteria. The comparison to the PNV proposed by RF as discussed in this study was intended as a validation of VERDE, but, as a by product, it supported the PNV proposed by RF with a blind, purely data-driven approach.

The flexibility of the data-driven approach of VERDE is very useful when coupling it to climate models. The approach used with VERDE allows it to be very easily adapted to the climate information available and to the scheme that each climate model uses for characterizing the land surface. An example (M. Zampieri & P. Lionello unpublished) is the offline coupling of VERDE with the RegCM model, where VERDE adopts the land-cover scheme of RegCM and allows the climate model to compute all atmosphere-land surface interactions that would be produced by LCTs in equilibrium with the model climate. This flexibility and this sort of application cannot be provided by models or approaches such as that described by RF or Koeppen's climate classification, which are based on a choice of variables describing the land cover, which are not linked to the dynamics of climate models.

5. CONCLUSION

In this paper, we have proposed a diagnostic model of LCTs, called VERDE, which is based on cluster analysis of high-resolution datasets of observed vegetation distribution and climate. We have shown that VERDE is a useful tool for computing realistic LCTs in equilibrium with climate and PNV.

VERDE is flexible, in that it can be used with any standard vegetation or land-cover classification, is user-friendly (its implementation is relatively easy), and is inexpensive (running VERDE takes a few seconds on a common desktop computer). It can be coupled offline to climate models, and the new LCTs computed by VERDE on the basis of the computed climate could be immediately used by the models themselves in subsequent simulations (M. Zampieri & P. Lionello unpublished). Moreover, VERDE could be improved by introducing other climatic and non-climatic variables, such as soil type, which clearly affects vegetation types (e.g. Bachelet et al. 1998), because of the important interplay of climate with the soil's capability to retain water and make it available for plant growth.

We used VERDE to reconstruct vegetation distribution in some areas (such as Europe, India, and China) where vegetation has been replaced because of anthropic activities. The dominant PNV would be broadleaf deciduous trees in Central Europe, with a transition to mixed savanna and grassland in the east at mid-latitudes, and evergreen needle trees in Russia. Large areas would be savanna in India, and grassland, mixed forest, and evergreen broadleaf trees in China. The results of VERDE compare reasonably well with previous PNV reconstructions (e.g. Ramankutty & Foley 1999).

VERDE was applied to 5 climate model A2 scenario simulations in order to identify the LCT changes in critical areas as a consequence of projected climate change. In the Northern Hemisphere, our results showed an increase in barren soils (deserts) within the tropics and at mid-latitudes, a northward shift of forests, and a reduction of shrublands and snow-ice covered soil at high latitudes. Changes were smaller and less consistent among different models in the Southern Hemisphere.

Acknowledgements. M.Z. was supported by the CLIMESCO project. The k-means cluster analysis was performed using the free statistical software R.

LITERATURE CITED

➤ Alo CA, Wang G (2008) Hydrological impact of the potential future vegetation response to climate changes projected by 8 GCMs. J Geophys Res 113:G03011. doi:10.1029/2007 JG000598

➤ Bachelet D, Brugnach M, Neilson RP (1998) Sensitivity of a biogeography model to soil properties. Ecol Model 109: 77–98

➤ Bond WJ, Woodward FI, Midgley GF (2004) The global distribution of ecosystems in a world without fire. New Phytol 165:525–538

➤ Cohen J (1960) A coefficient of agreement for nominal scale. Educ Psychol Meas 20:37–46

➤ Cramer W, Bondeau A, Woodward FI, Prentice C and others (2001) Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models. Glob Change Biol 7: 357–373

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell PAMI-1:224–227

Härdtle W (1995) On the theoretical concept of the potential natural vegetation and proposals for an up-to-date modification. Folia Geobot 30:263–276

Haxeltine A, Prentice IC (1996) BIOME3: an equilibrium terrestrial biosphere model based on ecophysiological constraints, resource availability, and competition among plant functional types. Global Biogeochem Cycles 10:693–709

Köppen W (1900) Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt. Geogr Z 6:593–611, 657–679

Kottek M, Grieser J, Beck C, Rudolf B, Rubel F (2006) World map of the Köppen-Geiger climate classification updated. Meteorol Z 15:259–263

Krinner G, Viovy N, de Noblet-Ducoudré N, Ogée J and others (2005) A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. Glob Biogeochem Cycles 19:GB1015

Lapola DM, Oyama MD, Nobre CA, Sampaio G (2008). A new world natural vegetation map for global change studies. An Acad Bras Ciênc 80:397–408 doi: 10.1590/S0001-3765 2008000200017

Loveland TR, Reed BC, Brown JF, Ohlen DO, Zhu J, Yang L, Merchant JW (2000) Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. Int J Remote Sens 21:1303–1330

MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proc 5th Berkeley Symp Math Stat Probability. University of California Press, Berkeley, CA, p 281–297

Matthews E (1983) Global vegetation and land use: new high-resolution data bases for climate studies. J Clim Appl Meteorol 22:474–487

Moravec J (1998) Reconstructed natural versus potential natural vegetation in vegetation mapping: a discussion of concepts. Appl Veg Sci 1:173–176

Nakićenović NJ, Swart R (eds) (2000) IPCC special report on emissions scenarios. Cambridge University Press, Cambridge

Neuhausl R (1991) Vegetation map of Europe: first results and current state. J Veg Sci 2:131–134

New M, Lister D, Hulme M, Makin I (2002) A high-resolution data set of surface climate over global land areas. Clim Res 21:1–25

Peel MC, Finlayson BL, McMahon TA (2007) Updated world map of the Köppen-Geiger climate classification. Hydrol Earth Syst Sci 11:1633–1644

Ramankutty N, Foley JA (1999) Estimating historical changes in global land cover: croplands from 1700 to 1992. Global Biogeochem Cycles 13:997–1027

Rosati L, Marignani M, Blasi C (2008) A gap analysis comparing Natura 2000 vs National Protected Area network with potential natural vegetation. Community Ecol 9: 147–154

Sitch S, Smith B, Prentice IC, Arneth A and others (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ Dynamic Vegetation Model. Glob Change Biol 9:161–185

Tüxen R (1956) Die heutige potentielle natürliche Vegetation als Gegenstand der Vegetationskartierung. Angew Pflanzensoziol (Stolzenau) 13:5–42

Unal Y, Kindap T, Karaca M (2003) Redefining the climate zones of Turkey using cluster analysis. Int J Climatol 23: 1045–1055

Wang A, Price DT (2007) Estimating global distribution of boreal, temperate, and tropical tree plant functional types using clustering techniques. J Geophys Res 112:G01024. doi:10.1029/2006JG000252

Woodward FI, Smith TM, Emanuel WR (1995) A global land primary productivity and phytogeography model. Global Biogeochem Cycles 9:471–490

Yang ZL, Dickinson RE (1996) Description of the Biosphere-Atmosphere Transfer Scheme (BATS) for the Soil Moisture Workshop and evaluation of its performance. Global Planet Change 13:117–134

Appendix 1. Cluster algorithm

In this study, the k-means cluster algorithm was applied to a climate and a land cover type (LCT) dataset, represented with the vector $\tilde{x} = (\tilde{t}, \tilde{p}, \tilde{v})$, where *t*, *P*, *v* are temperature, precipitation and LCT, as in Eqs. (1–3). Clusters are computed minimizing the total distance $\tilde{D}$, defined as the sum of the distances of the cluster elements $\boldsymbol{C}_{il}$ from the respective centroid $\boldsymbol{C}_l$:

$$\tilde{D}^2 = \sum_{l=1}^{N_C} \sum_{i=1}^{N_l} dist^2(\tilde{x}_{il}, \tilde{x}_{C_l}) \qquad \text{(A1)}$$

where N_C is the number of clusters and N_l is the number of elements on the *l*th cluster, and the distance dist is given by the usual Euclidean definition in an N_D-dimensional space.

$$\text{dist}^2(\tilde{x}, \tilde{y}) = \sum_{p=1}^{N_D} (\tilde{x}_p - \tilde{y}_p)^2 \qquad \text{(A2)}$$

where in this study the index *P* denotes the 24 + $\boldsymbol{N_V}$ components of the vector *x*.

A substantial amount of work has been used for the identification of the optimal number of clusters. Fig. A1a shows the behavior of the total distance $\tilde{D}$ as a function of the number of clusters. Note that $\tilde{D}$ is the sum of 3 contributions:

$$\tilde{D}^2 = \tilde{D}_T^2 + \tilde{D}_P^2 + \tilde{D}_V^2 \qquad \text{(A3)}$$

where

$$\tilde{D}_T^2 = \sum_{l=1}^{N_C} \sum_{i=1}^{N_l} \tilde{d}_{T_{il}}^2, \quad \tilde{d}_{T_{il}}^2 = \sum_{m=1}^{12} (\tilde{t}_{il}^{(m)} - \tilde{t}_{C_l}^{(m)})^2 \qquad \text{(A4)}$$

and similar definitions hold for $\tilde{D}_P^2$ and $\tilde{D}_V^2$, which describe the distances in the subspaces scanned varying the $t(m)$, $p(m)$, and $v(m)$ coordinates. $\tilde{D}$ necessarily diminishes as the number of clusters increases. Fig. A1 shows that the minimization proceeds evenly and smoothly for all 3 contributions, so that each corresponds to approximately one-third of the total distance, regardless of the number of clusters. Fig. 7b shows the average (SD) within clusters for the dimensional variables $t(m)$, $p(m)$, and $v(m)$, where

$$\text{SD}_{\text{int}\,ra}^2(T) = \frac{1}{N_c} \sum_{l=1}^{N_C} \frac{1}{N_l} \sum_{i=1}^{N_l} \frac{1}{12} \sum_{m=1}^{12} (\tilde{t}_{il}^{(m)} - \tilde{t}_{C_l}^{(m)})^2 \qquad \text{(A5)}$$

and similar definitions are used for $\text{SD}_{\text{intra}}(P)$ and $\text{SD}_{\text{intra}}(V)$. This plot shows, in dimensional units (K, mm) and percentage, the average deviations of a single element of the clus-

Appendix 1 (continued)

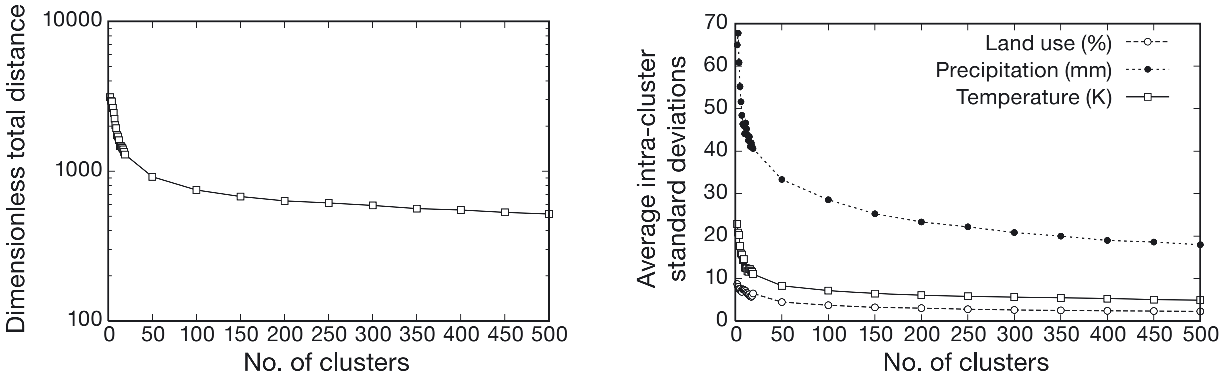

Fig. A1. (a) Total distance D as a function of the number of clusters; (b) average intra-cluster SD for temperature (K), precipitation (mm), and vegetation (land use, %). Both panels show these quantities as a function of the number of clusters

ter from its centroid. As the number of clusters increases and the number of elements within each cluster decreases, both average deviation and total distance diminish.

A criterion for identifying a suitable number of clusters is the inter-intra distance ratio R (Davies & Bouldin 1979):

$$R^2 = \frac{\tilde{\delta}^2_{\text{inter}}}{\tilde{\delta}^2_{\text{intra}}} \qquad \text{(A6)}$$

The quantity R can be computed as function of the number of clusters, where $\tilde{\delta}_{\text{inter}}$ is the average distance among different clusters, defined using the cluster centroids

$$\tilde{\delta}^2_{\text{inter}} = \frac{1}{N_C(N_C-1)}\sum_{m>l} dist^2(\tilde{x}_{C_l},\tilde{x}_{C_m}) \qquad \text{(A7)}$$

and $\tilde{\delta}_{\text{intra}}$ is the average distance within each cluster defined considering the distance between cluster elements and the cluster centroid

$$\tilde{\delta}^2_{\text{intra}} = \frac{1}{N_C}\sum_{l=1}^{N_C}\frac{1}{N_L}\sum_{i=1}^{N_L} dist^2(\tilde{x}_{il},\tilde{x}_{C_l}) \qquad \text{(A8)}$$

A large value of R denotes a situation in which the cluster centroids are well separated with respect to the dispersion of the cluster elements. In an ideal situation, all elements are grouped in well separated small clouds centered around the cluster centroids. The inter-intra distance ratio R is shown as a function of the number of clusters in Fig. A2. The figure also shows the values R_T, R_P, and R_V computed exactly as R, but using only variables in the subspaces scanned by $t(m)$, $p(m)$, and $v(m)$, respectively. Fig. A2 shows that the growth of R, R_T, R_P, and R_V slows down as the number of cluster increases, but it does not stop. The steepness of these 4 curves when the cluster number is comparatively low suggests that it would not be advisable to use less than about 100 clusters, but it fails to provide a clear indication of the upper limit beyond which it is not worth increasing their number. In this study, all applications were performed using 100 clusters, as described in Section 2.

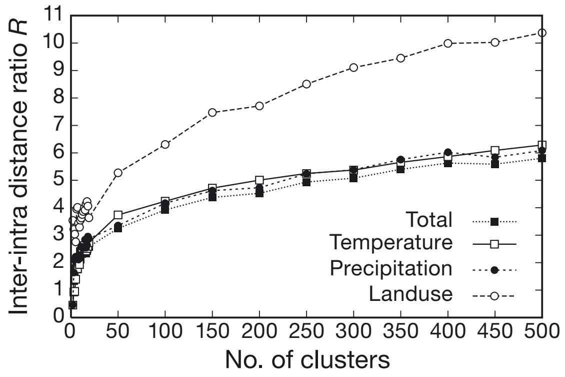

Fig. A2. Inter-intra distance ratio R (indicated in the key as 'Total') as a function of the number of clusters together with values of R for temperature (R_T), precipitation (R_P) and vegetation (landuse, R_V)

When the number of clusters is large, the clustering algorithm produces clusters with very few elements, meaning that they can be attributed to small areas (few grid points). The meaning and robustness of these small clusters is arguable. They might represent a local scale situation in which vegetation is not in equilibrium with the climate, such as remnants of vegetation types that are going extinct because they are no longer adapted, or they might be due to local factors, such as special soil conditions, that do not exist anywhere else at a global scale, or to human intervention. In contrast, the large size of a cluster is an indication of a robust link between climate and vegetation. Moreover, a series of tests that we carried out while developing the model has shown that small clusters produce instability during the attribution, as clusters including very few elements according to the analysis (first step of the procedure) might be attributed (second step of the procedure) to large areas, and this attribution depends irregularly on the number of clusters. This problem is avoided if small clusters are not used during the second step (attribution) of the procedure. The definition of small depends on the total number of clusters, because as the number of clusters increases, the average cluster size decreases. Therefore, a dimensionless size is defined as the ratio between the number of elements in each cluster and the average number of elements per cluster. Fig. A3 shows the fraction

Appendix 1 (continued)

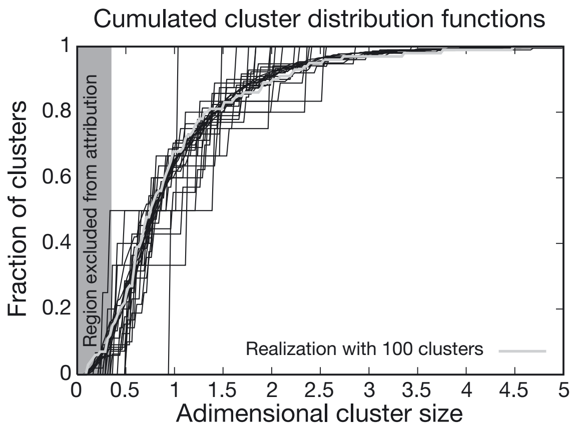

Fig. A3. Fraction of clusters (y-axis) exceeding a fixed dimensionless size (x-axis). Each line refers to a computation based on a different number of clusters (from 2 to 500). The line corresponding to the implementation with 100 clusters is marked light grey

of clusters exceeding a given dimensionless size. It shows that 10% of the clusters are smaller than 0.4 the average size independently from the total number of clusters used by the model. Empirical tests that were run while developing VERDE have shown that if clusters smaller than 0.4 times the average size are not considered when attributing a cluster to a model grid point, the robustness of the results greatly increases (see Section 2).

With reference to the discussion of VERDE validation over the Americas (final part of Section 2) and the relative maps, the distance in the subspace scanned by $\tilde{t}_{ij}^{(m)}$ and $\tilde{p}_{ij}^{(m)}$ between each point ij and the centroid C_{ij} of the cluster containing the grid point itself (Fig. 3d) is defined as:

$$\mathrm{dist}^2_{TP_{ij}} = \mathrm{dist}^2_{T_{ij}} + \mathrm{dist}^2_{P_{ij}} = \sum_{m=1}^{12} (\tilde{t}_{ij}^{(m)} - \tilde{t}_{C_{ij}}^{(m)})^2 + \sum_{m=1}^{12} (\tilde{p}_{ij}^{(m)} - \tilde{p}_{C_{ij}}^{(m)})^2 \quad (\mathrm{A9})$$

Fig. 3e shows a dimensional quantity (values in K), that is the deviation around the centroid

$$\mathrm{SD}^2_T = \frac{1}{12}\sum_{m=1}^{12} (t_{ij}^{(m)} - t_{C_{ij}}^{(m)})^2 \quad (\mathrm{A10})$$

Fig. 3f shows the analogous quantity for precipitation (values in mm).

Appendix 2. Land cover type (LCT) redefinition

This appendix shows the scheme used for comparing the results produced by VERDE, which adopts the GLCC IGBP LCTs for building the clusters, to the PNV reconstructed by Ramankutty & Foley (1999, hereafter referred to as 'RF'). The scheme consists of rules for converting the VERDE LCTs to the vegetation types used by RF. The purpose was to obtain global maps using the same classification for describing the land cover.

The GLCC IGBP LCTs consist of 17 types: (1) evergreen needle forest, (2) evergreen broadleaf forest, (3) deciduous needle forest, (4) deciduous broadleaf forest, (5) mixed forest, (6) closed shrublands, (7) open shrublands, (8) woody savannas, (9) savannas, (10) grasslands, (11) permanent wetlands, (12) croplands, (13) urban and built-up, (14) cropland/natural vegetation mosaic, (15) snow and ice, (16) barren or sparsely vegetated, (17) water bodies. of these types, 5 (the 4 'non natural' types, 11–14, and (17), were not considered in this study. These types were compared to the 16 types used by RF: (1) tropical evergreen forest/woodland, (2) tropical deciduous forest/woodland, (3) temperate broadleaf evergreen forest/woodland, (4) temperate needle evergreen forest/woodland, (5) temperate deciduous forest/woodland, (6) boreal evergreen forest/woodland, (7) boreal deciduous forest/woodland, (8) mixed forest, (9) savanna, (10) grassland/steppe, (11) dense shrubland, (12) open shrubland, (13) tundra, (14) desert, (15) polar desert/rock/ice.

The comparison in Fig. 5 was carried out adopting the rules below.

A few rules are obvious. When the dominant vegetation of the cluster centroid is (5), (6), or (9) (according to the GLCC IGBP classification), the following associations are adopted:

- (5) GLCC IGBP → (8) RF
- (6) GLCC IGBP → (11) RF
- (9) GLCC IGBP → (9) RF
- (10) GLCC IGBP → (10) RF

Besides the dominant LCT of the cluster centroid, other rules also consider the monthly mean temperature values and accumulated precipitation of the cluster centroid and are guided by the usual definition of climate types (e.g. Köppen):

- (1) and (2) GLCC IGBP → (1) RF, if mean monthly temperature is >15°C and monthly precipitation exceeds 1500 mm
- (1) and (2) GLCC IGBP → (6) RF if mean annual temperature is <0°C or the month has a mean temperature >10°C → (4) RF and (3) RF, respectively, in all other cases
- (3) and (4) GLCC IGBP → (2) RF if mean monthly temperature is >15°C and monthly precipitation exceeds 1500 mm
- (3) and (4) GLCC IGBP → (7) RF if mean annual temperature is <0°C no month has a mean temperature >10°C
- (3) and (4) GLCC IGBP → (5) RF otherwise
- (16) GLCC IGBP → (13) RF if the cluster has no month with mean temperature >10°C; → (14) RF otherwise
- (7) GLCC IGBP → (13) RF if mean annual temperature <0°C; → (12) RF otherwise.

Other rules consider the LCT with the second largest percentage in the cluster centroid.

- (8) GLCC IGBP → (9) RF if the second largest LCT is savanna; → (1) RF If the second largest LCT is evergreen forest; → (2) RF If second largest LCT is deciduous forest
- (15) GLCC IGBP → (13) RF if a second vegetation type is present with a fraction higher than 0.3 → (15) RF in all other cases

Editorial responsibility: Gerd Esser,
Gießen, Germany

Submitted: July 9, 2009; Accepted: January 26, 2010
Proofs received from author(s): April 18, 2010