# Statistical modelling of a new global potential vegetation distribution

OPEN ACCESS
IOP Publishing

Environ. Res. Lett. **7** (2012) 044019 (11pp)
doi:10.1088/1748-9326/7/4/044019

# Statistical modelling of a new global potential vegetation distribution

**G Levavasseur[1], M Vrac[1], D M Roche[1,2] and D Paillard[1]**

[1] Laboratoire des Sciences du Climat et de l'Environnement (LSCE), UMR 8212, IPSL—CEA/CNRS-INSU/UVSQ, Centre d'étude de Saclay, Orme des Merisiers, F-91191 Gif-sur-Yvette, France
[2] Section Climate Change and Landscape Dynamics, Department of Earth Sciences, Faculty of Earth and Life Sciences, VU University Amsterdam, de Boelelaan 1085, NL-1081 HV Amsterdam, The Netherlands

E-mail: Guillaume.Levavasseur@lsce.ipsl.fr

Received 30 May 2012
Accepted for publication 12 October 2012
Published 31 October 2012
Online at stacks.iop.org/ERL/7/044019

**Abstract**
The potential natural vegetation (PNV) distribution is required for several studies in environmental sciences. Most of the available databases are quite subjective or depend on vegetation models. We have built a new high-resolution world-wide PNV map using a objective statistical methodology based on multinomial logistic models. Our method appears as a fast and robust alternative in vegetation modelling, independent of any vegetation model. In comparison with other databases, our method provides a realistic PNV distribution in agreement with respect to BIOME 6000 data. Among several advantages, the use of probabilities allows us to estimate the uncertainty, bringing some confidence in the modelled PNV, or to highlight the regions needing some data to improve the PNV modelling. Despite our PNV map being highly dependent on the distribution of data points, it is easily updatable as soon as additional data are available and provides very useful additional information for further applications.

**Keywords:** statistical model, potential vegetation, multinomial logistic regression, biomes, world, climate

S Online supplementary data available from stacks.iop.org/ERL/7/044019/mmedia

## 1. Introduction

The 'potential natural vegetation' (PNV) can be seen as the natural vegetation, i.e., in equilibrium with climate, that would exist at a given location non-impacted by human activities. A global PNV distribution is required for many purposes in environmental sciences. Examples include estimating historical changes of land-use (Ramankutty and Foley 1999), the impact of atmospheric $CO_2$ concentration on vegetation (Cha 1997, Notaro *et al* 2005), the response of vegetation to climate changes (Ni *et al* 2006, Notaro 2008) or palaeovegetation distributions (Crucifix *et al* 2005, Woillez *et al* 2011).

Most of these studies refer to the global PNV map derived by Ramankutty and Foley (1999) (hereafter 'RF99') from remotely-sensed observations and corrected in human-impacted regions using the vegetation model BIOME3 (Haxeltine and Prentice 1996). Other applications (e.g., palaeovegetation modelling) directly use vegetation models to simulate a PNV distribution. These two methods include biases and uncertainties from vegetation models.

In this context, Levavasseur *et al* (2012) (hereafter 'L12') described a new methodology to statistically model a high-resolution PNV distribution over Europe, entirely based on vegetation and climatological data. Their approach consists in using a multinomial logistic regression (MLR). MLR builds statistical relationships between vegetation data and climatological variables independently from models and without any subjective corrections. These relationships allow one to model the occurrence probabilities of each PNV type. L12 show good results over Europe in comparison with the map from RF99 or the PNV simulated by the vegetation model BIOME4 (Kaplan *et al* 2003). Moreover, the use of occurrence probabilities provides useful additional information as PNV fractions or uncertainty index.

In the present work we extend the L12 method to model a high-resolution gridded world-wide PNV distribution. We summarize the L12 framework and highlight the adaptations and updates that we performed in section 2. Then, we compare the global PNV modelled by MLR to the RF99 map (section 3.3) and to the PNV distribution simulated by the vegetation model BIOME4 over the globe (section 3.2). Discussions and conclusions follow in section 4.

## 2. Method

### 2.1. The multinomial logistic regression (MLR)

To predict PNV types distribution, L12 used a multinomial logistic regression (MLR, Hosmer and Lemeshow 2000, Hilbe 2009). MLR builds statistical relationships between a nominal explained variable (called the predictand, the vegetation type in our case) and continuous explanatory variables (called the predictors, the climatic variables). Those relationships allow one to estimate the occurrence probability of the nominal explained variable ($Y$, the PNV types in our case), taking into account $p$ numerical explanatory variables ($X_k$, the climatological variables):

$$\log\left(\frac{P(Y_i=j)}{P(Y_i=r)}\right) = \beta_{0,j} + \sum_{k=1}^{p} \beta_{k,j} X^*_{i,k}, \qquad \forall j \neq r, \quad (1)$$

where $X^*_{i,k} = \frac{X_{i,k}-\mu_k}{\sigma_k}$ are the standardized predictors, with $\mu_k$ the mean of the $k$th predictor and $\sigma_k$ its standard deviation. $P(Y_i = j)$ is the occurrence probability of the $j$th PNV type and $i$ is the grid-cell. $\beta_{0,j}$ is the intercept for the $j$th PNV type and $\beta_{k,j}$ is the regression coefficient for the $k$th predictor and the $j$th PNV type. $p$ is the number of predictors.

The predictors are standardized to obtain comparable regression coefficients ($\beta_{k,j}$) without units. According to their weights, the predictors will be ranked and selected in section 2.4.

Equation (1) is based on a reference category $r$: the desert vegetation in our case (defined in section 2.2.2). We obtain $j - 1$ relationships and the occurrence probabilities of the reference PNV type can be deduced in each grid-cell $i$ with $\sum_{j=1}^{m} P(Y_i = j) = 1$ (considering $m$ PNV types including $r$). MLR is performed with the R package 'VGAM' (Yee and Wild 1996, Yee 2010a, 2010b) and parameters are estimated through likelihood maximization.

### 2.2. Vegetation data

*2.2.1. The BIOME 6000 database.* As predictand for MLR, we use the BIOME 6000 database (Prentice and Jolly 2000, Harrison *et al* 2001, Bigelow *et al* 2003, Pickett *et al* 2004) from the Global Palaeovegetation Mapping Project[3] for the modern period (i.e., 0 ka). BIOME 6000 compiles pollen and plant macrofossil data that are heterogeneously distributed over the world. Most data points are concentrated in the northern hemisphere between 30°N and 70°N, and no data point covers South America or India. Our region of interest is the globe without Antarctica, from 180°W to 180°E and from 60°S to 90°N, as shown in figure 1(a).

BIOME 6000 data points are expressed in biomes following a 'biomization' method described in Prentice *et al* (1996). A biome includes characteristic vegetation types deduced from pollen under similar climatic conditions. BIOME 6000 can be classified into eight 'megabiomes': boreal forest, desert, tundra, grassland and dry shrubland, savanna and dry woodland, temperate forest, tropical forest and warm-temperate forest.

According to the authors of the database, some BIOME 6000 data points appear inconsistent in mountain areas due to pollen transport (Guiot 2012). For instance, the Alps or Pyrenees are essentially dominated by temperate forest pollens, even at high altitude. To correct for such discrepancies, we reclassify some BIOME 6000 data points using the growing degree day at 5 °C (GDD5) limits from Prentice *et al* (1992) (see figure 1(a)). Annual GDD5 corresponds to the sum of daily temperatures above 5 °C during a year. As no gridded dataset of daily temperature covers the world with a fine enough spatial resolution, we first compute a global GDD5 climatology from the NCEP/NCAR[4] air temperature at the surface reanalysis daily time-series (between 1961 and 1990) on a 2.5° grid (Kalnay *et al* 1996). Then, we apply a statistical downscaling method based on the use of a Generalized Additive Model (GAM, Vrac *et al* 2007) to obtain a global GDD5 (cf appendix A) at a resolution of 10′ (i.e., 1/6° in longitude and latitude, the final resolution of our map). This procedure reclassified 221 BIOME 6000 data points (i.e., 3.6% of the 6091 points) essentially localized on mountains (Alps, Himalaya, Rocky Mountains) and in boreal regions (Alaska, the coast of Labrador and Siberia).

*2.2.2. 'True' deserts.* We define true deserts as sterile land areas without vegetation producing pollens. Consequently, no BIOME 6000 data point covers desert regions such as the Sahara or the Greenland ice-sheet. The desert megabiome from BIOME 6000 refers more specifically to desert vegetation dominated by sparse steppe forb and grass (Prentice *et al* 1996) instead of 'true' desert. Accordingly,

[3] BIOME 6000 database version 4.2 available on www.bridge.bris.ac.uk/resources/Databases/BIOMES_data/.

[4] The National Centers for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR) reanalysis data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their web site at www.esrl.noaa.gov/psd/.

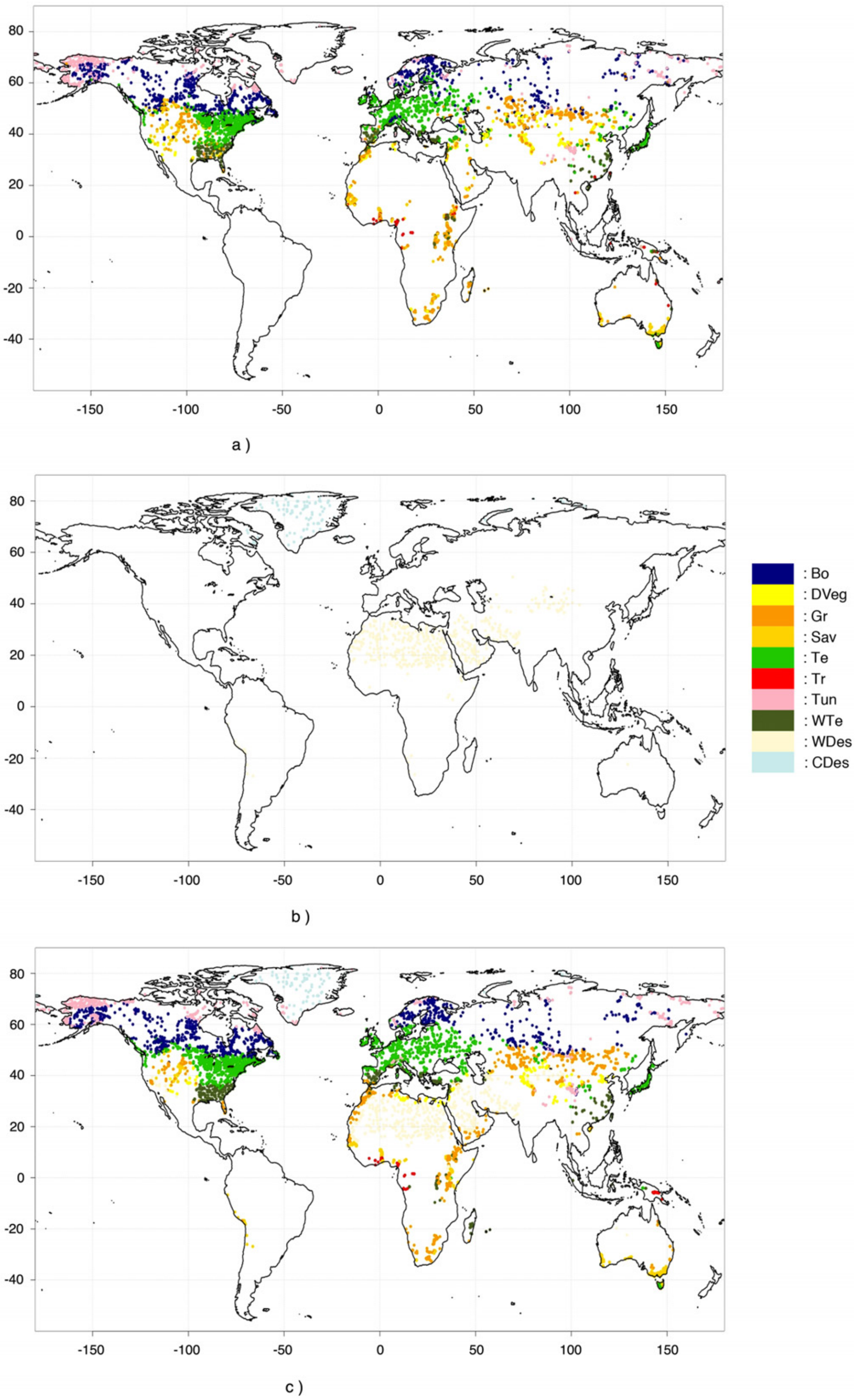

**Figure 1.** BIOME 6000 data (a), the added points for cold and warm deserts (b) (see section 2.2.1) and the PNV predicted by MLR in each training point (c). In legend, 'Bo' stands for boreal forests, 'DVeg' for desert vegetation, 'Gr' for grasslands and dry shrublands, 'Sav' for savannas and dry woodlands, 'Te' for temperate forests, 'Tr' for tropical forests, 'Tun' for tundra, 'WTe' for warm-temperate forests, 'WDes' for warm deserts and 'CDes' for cold deserts.

we rename the desert megabiome from BIOME 6000 into a separated desert vegetation megabiome.

In order to represent real deserts, we build two new 'pseudo-megabiomes' (warm and cold deserts) from the

**Table 1.** List of the 43 potential predictors. The predictors selected for logistic regressions in table 2 are in bold letters.

| No | Name | Abbreviation | Unit | Sources/references |
|---|---|---|---|---|
| **1** | **Winter air temperature at the surface** | **TEMP.DJF** | °C | CRU—New *et al* (2002) |
| **2** | **Spring air temperature at the surface** | **TEMP.MAM** | — | — |
| **3** | **Summer air temperature at the surface** | **TEMP.JJA** | — | — |
| **4** | **Fall air temperature at the surface** | **TEMP.SON** | — | — |
| 5 | Winter diurnal cycle temperature range | DTEMP.DJF | °C | CRU—New *et al* (2002) |
| **6** | **Spring diurnal cycle temperature range** | **DTEMP.MAM** | — | — |
| 7 | Summer diurnal cycle temperature range | DTEMP.JJA | — | — |
| **8** | **Fall diurnal cycle temperature range** | **DTEMP.SON** | — | — |
| **9** | **Winter total precipitation** | **PREC.DJF** | mm | CRU—New *et al* (2002) |
| 10 | Spring total precipitation | PREC.MAM | — | — |
| 11 | Summer total precipitation | PREC.JJA | — | — |
| 12 | Fall total precipitation | PREC.SON | — | — |
| 13 | Winter wet days frequency | WET.DJF | days | CRU—New *et al* (2002) |
| 14 | Spring wet days frequency | WET.MAM | — | — |
| 15 | Summer wet days frequency | WET.JJA | — | — |
| 16 | Fall wet days frequency | WET.SON | — | — |
| **17** | **Winter frost days frequency** | **FROST.DJF** | days | CRU—New *et al* (2002) |
| **18** | **Spring frost days frequency** | **FROST.MAM** | — | — |
| 19 | Summer frost days frequency | FROST.JJA | — | — |
| **20** | **Fall frost days frequency** | **FROST.SON** | — | — |
| 21 | Winter sunshine duration | SUN.DJF | % | CRU—New *et al* (2002) |
| 22 | Spring sunshine duration | SUN.MAM | — | — |
| 23 | Summer sunshine duration | SUN.JJA | — | — |
| **24** | **Fall sunshine duration** | **SUN.SON** | — | — |
| 25 | Winter relative humidity | RH.DJF | % | CRU—New *et al* (2002) |
| 26 | Spring relative humidity | RH.MAM | — | — |
| 27 | Summer relative humidity | RH.JJA | — | — |
| **28** | **Fall relative humidity** | **RH.SON** | — | — |
| 29 | Winter wind speed/intensity at 10 m | WND.DJF | m s$^{-1}$ | CRU—New *et al* (2002) |
| **30** | **Spring wind speed/intensity at 10 m** | **WND.MAM** | — | — |
| 31 | Summer wind speed/intensity at 10 m | WND.JJA | — | — |
| 32 | Fall wind speed/intensity at 10 m | WND.SON | — | — |
| 33 | Winter total cloudiness | CLOUD.DJF | % | CRU—New *et al* (2002) |
| 34 | Spring total cloudiness | CLOUD.MAM | — | — |
| 35 | Summer total cloudiness | CLOUD.JJA | — | — |
| 36 | Fall total cloudiness | CLOUD.SON | — | — |
| **37** | **Winter growing degree day at 5 °C** | **GDD.DJF** | °C days | NCEP/NCAR—Kalnay *et al* (1996) |
| **38** | **Spring growing degree day at 5 °C** | **GDD.MAM** | — | — |
| **39** | **Summer growing degree day at 5 °C** | **GDD.JJA** | — | — |
| 40 | Fall growing degree day at 5 °C | GDD.SON | — | — |
| 41 | Topography | TOPO | m | NGDC—Amante and Eakins (2009) |
| 42 | East–west topographic gradient | GTEW | % | — |
| 43 | North–south topographic gradient | GTNS | — | — |

IGBP-DIS land cover map[5] (Loveland and Belward 1997). This dataset derived 17 land cover types from remotely-sensed observations between 1992 and 1993. Assuming a limited human-induced desertification, we manually add training data points as follows (see figure 1(b)):

- To reflect cold deserts, 100 points have been randomly evenly distributed over the Greenland ice-sheet and where Loveland and Belward (1997) shows polar, rock or ice deserts. 100 points are enough to homogeneously cover cold desert regions.
- Over the globe, the total warm desert area is approximately five times larger than the total cold desert area. To be consistent and reflect warm deserts, 500 data points have been distributed in the same way as where Loveland and Belward (1997) shows warm deserts.

### *2.3. The explanatory variables*

We deal with the same climatological and geographical variables as used in L12 but with the downscaled GDD5 described in section 2.2.1 and appendix A. Table 1 lists the 43 potential predictors divided into two groups, the ‘climatic’ predictors and the ‘geographical’ ones.

[5] International Geosphere–Biosphere Programme Data and Information System, data set available on-line: http://daac.ornl.gov/.

- Climatological variables are taken from the Climate Research Unit (CRU) database[6] (New *et al* 2002) available at a regular spatial resolution of 10′. For each grid-point the dataset counts twelve monthly means (from 1961 to 1990) and each variable is divided into four 'seasonal' predictors by averaging data over the three corresponding months (e.g., 'TEMP.DJF' stands for winter temperature).
- Geographical variables are computed from the high-resolution gridded dataset ETOPO[7] at 10′ resolution (Amante and Eakins 2009) from the National Geophysical Data Center (NGDC).

### 2.4. Model selection

To avoid modelling any vegetation in a desert region and interfering with the model selection, we run three logistic regressions:

(i) Cold deserts are modelled by a first binary logistic regression. The explained variable is a binary vector indicating whether the data point is a cold desert or not.

(ii) Warm deserts are modelled by a second binary logistic regression in each grid-cell without cold deserts. The explained variable is a binary vector indicating whether the data point is a warm desert or not.

(iii) Finally, a multinomial logistic regression models the eight megabiomes for each grid-cell with no deserts. For this step, the explained variables are the BIOME 6000 data points.

Taking into account the 43 predictors (table 1) leads to an excessively complex statistical model, reducing its predictive performance by over-fitting. Moreover, a high correlation could exist between predictors, providing redundant information (Levavasseur *et al* 2012). To avert these issues, we select the model with the most appropriate combination of predictors. It would be too computationally intensive to test all possible combinations of predictors (i.e., $2^{43}$) for each logistic regression. Therefore, we use the following procedure:

(i) We run a calibration with all 43 standardized predictors ($X^*_{i,k}$ in equation (1)) for each logistic regression. We select predictors carrying more than 5% of the overall information/variability for each megabiome (which could be different depending on the megabiome) according to their regression coefficients ($\beta_{k,j}$ in equation (1)): five predictors for warm deserts, four for cold deserts and 15 predictors for the eight megabiomes.

(ii) Each possible combination among the pre-selected predictors has been tested, plus the 'null-model' corresponding to a model with only the intercepts ($\beta_{0,j}$ in equation (1), i.e., all regression coefficients $\beta_{k,j}$ are 0).

(iii) For each logistic regression, we select the best predictors set according to the Bayesian Information Criterion (BIC) described in appendix B. This index balances between the goodness-of-fit and the complexity (i.e., the number of parameters and predictors) of the tested model.

## 3. Results

### 3.1. Comparison MLR versus BIOME 6000

Table 2 summarizes the best model, i.e., with the smallest BIC, for each logistic regression. MLR models the occurrence probability of each vegetation type. For each grid-cell with no warm or cold desert modelled by the two binary logistic regressions, we take the megabiome with the maximum occurrence probability modelled by the third logistic regression as the dominant megabiome.

Figure 1(c) shows the predicted megabiomes by MLR in each training point location. In comparison with figures 1(a) and (b), the PNV modelled by MLR locally differs where BIOME 6000 shows several megabiomes at the same location or under-represents a megabiome in a region. For instance, MLR models grasslands and dry shrublands in the east of the Caspian Sea instead of desert vegetation in BIOME 6000; it replaces the boreal forests of the US Rocky Mountains by savanna or grasslands. The climatic signal provided by the predictors could be another cause of differences between both maps. Added desert points in the north and west of the Sahara are respectively replaced by desert vegetation and grasslands with MLR because of a fall relative humidity significantly lower in these regions (not shown).

Nevertheless, we note a good agreement between maps: 69.5% of BIOME 6000 data points (i.e., without the deserts) are correctly represented by MLR. Moreover, to quantify the quality of our modelling, we compute three other statistical indices excluding the added points for deserts: the kappa coefficient ($\kappa$), a pseudo-$R^2$ and the global Brier score (BS defined in appendix B). According to the classical scaling of the $R^2$ and the $\kappa$ coefficient used in vegetation studies (e.g., Monserud and Leemans 1992), a pseudo-$R^2$ of 0.57 and a $\kappa$ of 0.64 confirm a global good agreement with BIOME 6000 data. A BS of 0.41, far from 8 (the maximum value indicating bad agreement), attests the accuracy of the occurrence probabilities and of the PNV modelled by MLR.

### 3.2. Comparison MLR versus BIOME4

To ascertain our method, we directly confront the modelled PNV by MLR with the simulated vegetation from a vegetation model. The BIOME4 model (Haxeltine and Prentice 1996, Kaplan *et al* 2003) is driven by temperature, sunshine and precipitation monthly climatologies from the CRU database (described in section 2.3). To be consistent with the period represented by CRU climatologies (around 1980), the atmospheric $CO_2$ concentration is set for 360 ppm (Lüthi *et al* 2008). BIOME 4 has a biome scale easily translatable into our megabiomes following Harrison and Prentice (2003).

[6] CRU databases CL 2.0 and CL 1.0 for cloudiness are available on www.cru.uea.ac.uk/cru/data/hrg/.

[7] Computerized digital images and associated databases are available from the National Geophysical Data Center, National Oceanic and Atmospheric Administration, US Department of Commerce, www.ngdc.noaa.gov/.

**Table 2.** The selected predictors after all possible combinations for each logistic regression described in section 2.4: the binary logistic regression for cold deserts (column 'CDes'), the binary logistic regression for warm deserts (column 'WDes'), and the multinomial logistic regression for BIOME 6000 megabiomes (without the reference category 'DVeg'—last seven columns). For each megabiome, the predictors are ranked according to their regression coefficients with: their names (first line), their values (second line) and their weights in per cent (third line). The predictors and megabiomes abbreviations are respectively set from table 1 and the legend of figure 1.

| Predictors ranking | Added deserts megabiomes | | BIOME 6000 megabiomes | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | CDes | WDes | Bo | Gr | Sav | Te | Tr | Tun | WTe |
| 1 | TEMP.MAM<br>20.1<br>43.5% | TEMP.MAM<br>4.4<br>50.7% | GDD.MAM<br>6.9<br>16.2% | TEMP.SON<br>6.5<br>24.8% | GDD.JJA<br>9.2<br>17.5% | TEMP.MAM<br>9.7<br>18.5% | TEMP.DJF<br>74.0<br>32.3% | TEMP.SON<br>6.4<br>15.3% | TEMP.DJF<br>16.1<br>30.0% |
| 2 | TEMP.DJF<br>19.3<br>41.7% | RH.SON<br>1.8<br>21.0% | TEMP.MAM<br>5.1<br>12.1% | TEMP.DJF<br>4.0<br>15.5% | TEMP.DJF<br>8.6<br>16.2% | TEMP.JJA<br>9.6<br>18.2% | TEMP.MAM<br>51.3<br>22.4% | GDD.MAM<br>5.7<br>13.7% | TEMP.MAM<br>5.9<br>10.9% |
| 3 | FROST.MAM<br>4.9<br>10.7% | FROST.SON<br>1.3<br>15.4% | TEMP.JJA<br>5.1<br>11.9% | GDD.JJA<br>3.6<br>13.7% | TEMP.JJA<br>7.5<br>14.2% | GDD.JJA<br>8.7<br>16.5% | GDD.MAM<br>35.2<br>15.4% | FROST.SON<br>5.2<br>12.4% | TEMP.SON<br>5.6<br>10.4% |
| 4 | DTEMP.MAM<br>1.9<br>4.1% | GDD.DJF<br>1.1<br>12.9% | FROST.SON<br>5.0<br>11.8% | FROST.SON<br>2.1<br>8.1% | FROST.SON<br>6.9<br>13.1% | TEMP.DJF<br>5.5<br>10.5% | GDD.DJF<br>24.9<br>10.9% | GDD.DJF<br>4.1<br>9.8% | GDD.JJA<br>5.5<br>10.2% |
| 5 | | | SUN.SON<br>4.6<br>10.7% | GDD.DJF<br>1.9<br>7.3% | TEMP.SON<br>6.6<br>12.4% | GDD.MAM<br>5.2<br>9.9% | TEMP.SON<br>13.8<br>6.0% | SUN.SON<br>3.6<br>8.6% | GDD.DJF<br>4.6<br>8.6% |
| 6 | | | GDD.DJF<br>3.5<br>8.2% | SUN.SON<br>1.7<br>6.7% | TEMP.MAM<br>5.5<br>10.4% | SUN.SON<br>4.0<br>7.7% | FROST.SON<br>10.9<br>4.8% | TEMP.DJF<br>3.3<br>8.0% | GDD.MAM<br>4.3<br>8.0% |
| 7 | | | FROST.DJF<br>3.2<br>7.6% | PREC.DJF<br>1.7<br>6.4% | SUN.SON<br>3.0<br>5.6% | PREC.DJF<br>2.8<br>5.2% | TEMP.JJA<br>6.2<br>2.7% | GDD.JJA<br>3.2<br>7.7% | TEMP.JJA<br>3.5<br>6.5% |
| 8 | | | TEMP.SON<br>2.9<br>6.8% | FROST.DJF<br>1.4<br>5.4% | FROST.DJF<br>2.1<br>4.0% | FROST.DJF<br>1.9<br>3.6% | FROST.DJF<br>4.0<br>1.8% | TEMP.MAM<br>3.0<br>7.1% | SUN.SON<br>2.5<br>4.6% |
| 9 | | | PREC.DJF<br>2.4<br>5.6% | GDD.MAM<br>1.3<br>5.0% | GDD.MAM<br>1.5<br>2.8% | TEMP.SON<br>1.8<br>3.5% | SUN.SON<br>3.3<br>1.5% | FROST.DJF<br>2.5<br>6.1% | PREC.DJF<br>2.2<br>4.1% |
| 10 | | | GDD.JJA<br>2.3<br>5.4% | TEMP.MAM<br>0.8<br>3.2% | PREC.DJF<br>1.0<br>1.9% | FROST.SON<br>1.6<br>3.1% | GDD.JJA<br>3.0<br>1.3% | PREC.DJF<br>2.2<br>5.2% | FROST.SON<br>1.3<br>2.4% |
| 11 | | | TEMP.DJF<br>0.9<br>2.1% | TEMP.JJA<br>0.4<br>1.6% | WND.MAM<br>0.5<br>1.0% | GDD.DJF<br>1.1<br>2.1% | PREC.DJF<br>1.7<br>0.8% | TEMP.JJA<br>2.0<br>4.7% | FROST.DJF<br>1.1<br>2.1% |
| 12 | | | DTEMP.SON<br>0.5<br>1.1% | DTEMP.SON<br>0.4<br>1.5% | DTEMP.SON<br>0.4<br>0.7% | DTEMP.SON<br>0.4<br>0.8% | WND.MAM<br>0.7<br>0.3% | DTEMP.SON<br>0.5<br>1.2% | WND.MAM<br>0.9<br>1.7% |
| 13 | | | WND.MAM<br>0.2<br>0.5% | WND.MAM<br>0.2<br>0.8% | GDD.DJF<br>0.1<br>0.1% | WND.MAM<br>0.2<br>0.4% | DTEMP.SON<br>0.1<br>0.0% | WND.MAM<br>0.1<br>0.3% | DTEMP.SON<br>0.3<br>0.5% |

Figures 2(a) and (b) respectively show the modelled PNV by MLR in each grid-cell of our map (at 10′ resolution) and the simulated PNV distribution by BIOME4. Both maps show large similarities especially concerning the distribution of tundra, temperate and boreal forests at high latitudes. Note that our BIOME 4 simulation does not show warm-temperate forests around Mediterranean Sea, in southeastern China and USA and in eastern Australia. Modelling warm-temperate forests by MLR in these regions is in agreement with BIOME 6000 database (see figure 1(a) and Levavasseur *et al* 2012) and with older published BIOME simulations (Prentice *et al* 1992, 1996, Harrison and Prentice 2003, Tang *et al* 2009).

Nevertheless, some mountain areas are not well defined (e.g., tundra and boreal forests of Tian Mountains are replaced by desert vegetation) or even disappear (e.g., the Andes or US Rockies) with MLR. Differences appears in the western US, where MLR models a drier vegetation than BIOME4. Moreover, MLR models larger warm deserts than BIOME4.

Although both methods are based on CRU climatologies, BIOME4 computes some mechanistic processes (i.e., physiology, competitiveness or productivity) which may induce some of the difference.

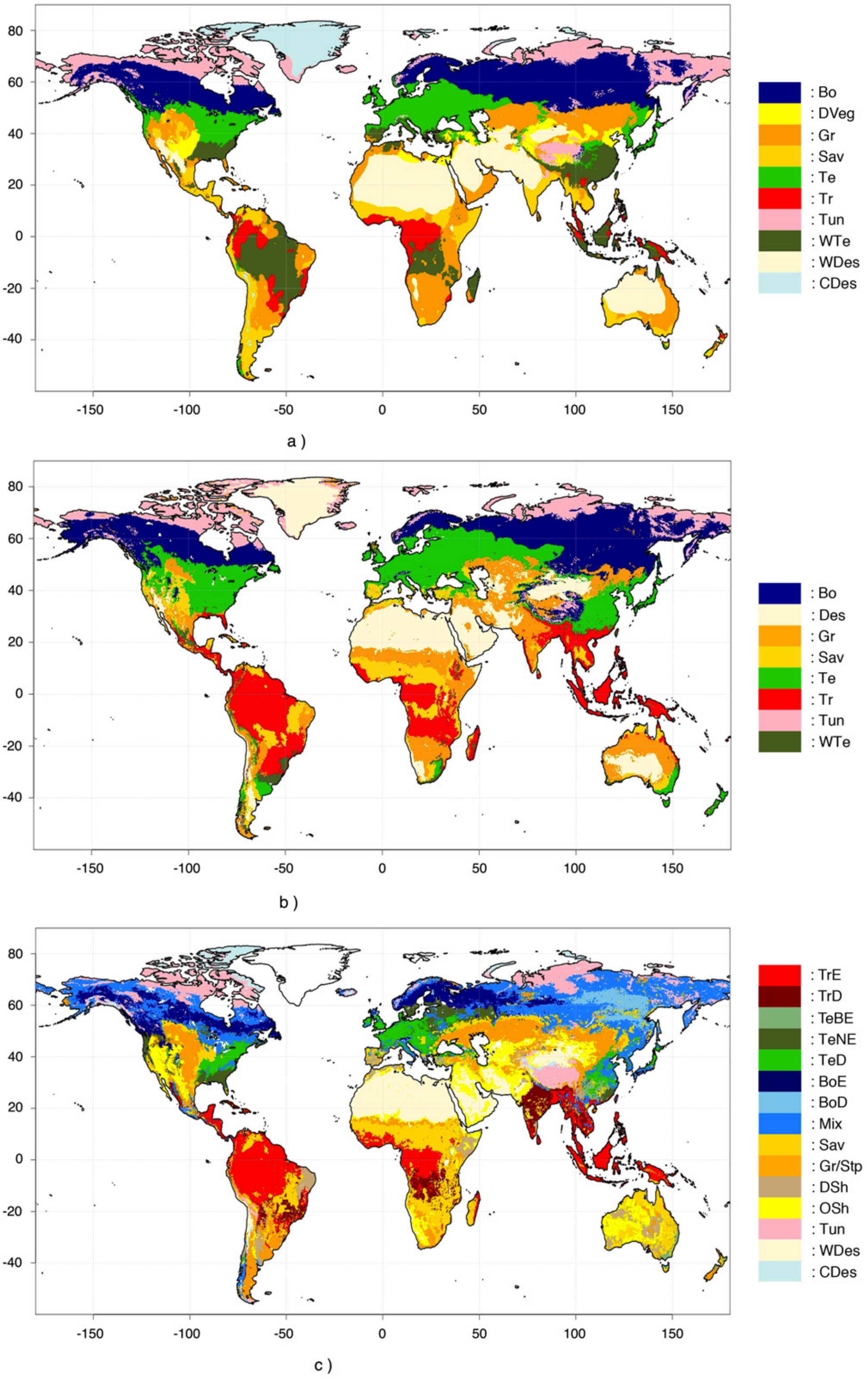

**Figure 2.** PNV distribution predicted by MLR (a) with the same biome scale as for figure 1(c). PNV distribution simulated by BIOME4 driven by CRU climatologies (b) with the same biome scale as for MLR grouping desert vegetation, warm desert and cold desert into the desert megabiome (‘Des’). The RF99 database (c) with its own vegetation types, where ‘TrE’ is set for tropical evergreen forests/woodlands, ‘TrD’ for tropical deciduous forests/Woodlands, ‘TeBE’ for temperate broadleaf evergreen forests/woodlands, ‘TeNE’ for temperate needleleaf evergreen forests/woodlands, ‘TeD’ for temperate deciduous forests/woodlands, ‘BoE’ for boreal evergreen forests/woodlands, ‘BoD’ for boreal deciduous forests/woodlands, ‘Mix’ for evergreen/deciduous mixed forests, ‘Sav’ for savannas, ‘Gr/Stp’ for grasslands/steppe, ‘DSh’ for dense shrublands, ‘OSh’ for open shrublands, ‘Tun’ for tundra, ‘WDes’ for warm deserts and ‘CDes’ for polar/rock/ice deserts.

### *3.3. Comparison MLR versus RF99*

In this section, we compare the modelled PNV distribution by MLR (figure 2(a)) to the RF99 map (figure 2(c)). Defining a correspondence between vegetation types from both databases highly depends on the region, e.g., mixed forests from RF99 have to be considered in BIOME 6000 as temperate forests in Europe and as boreal forest in Siberia (Levavasseur *et al* 2012). The PNV types from RF99 are deduced from real observed vegetation and are fundamentally different from a biome scale. Consequently, we choose to keep each map in its original scale (i.e., 9 megabiomes for MLR and 15 PNV types for RF99).

Both maps reveal a similar PNV distribution with tundra, temperate and boreal forests at high latitudes. In eastern Europe, the observed grasslands by RF99 are probably the result of deforestation (Kaplan *et al* 2009). Several global dynamical vegetation models simulate large areas of boreal forest in this region under preindustrial climatic conditions (Sitch *et al* 2003, Woillez *et al* 2011). The temperate forests modelled by MLR appear to be more likely in equilibrium with a warmer modern climate, as simulated by BIOME4 (figure 2(b)).

Nevertheless, MLR does not capture the impact of topography in the western US; RF99 shows a probably better distribution of shrublands, tundra and boreal forests in this mountain area. In equatorial regions, some tropical forests in RF99 are replaced by warm-temperate forests by MLR. The edges of warm deserts are also in disagreement depending on the region. RF99 sees tropical forests in India, while MLR modelled warm desert and savanna. In contrast, the Sahara does not reach the west and north African coast with MLR, which models grasslands and desert vegetation.

## 4. Conclusions and discussion

We compared in this paper several potential natural vegetation (PNV) distributions over the globe based on different methods:

- The PNV map from Ramankutty and Foley (1999) (RF99) built from remotely-sensed data (Loveland *et al* 2000) and from the vegetation model BIOME3 (Haxeltine and Prentice 1996).
- PNV simulated by the vegetation model BIOME4 (Kaplan *et al* 2003) driven by the CRU climatologies.
- A new high-resolution global PNV map built from multinomial logistic models.

Obvious similarities appear between reconstructions, especially with the establishment of tundra, temperate and boreal forests at high latitudes.

The vegetation model BIOME4 is partly calibrated to represent BIOME 6000 data (Kaplan *et al* 2003). The $\kappa$ coefficient computed between BIOME4 and BIOME 6000 (appendix B) is of 0.43 over the world, while MLR obtains a $\kappa$ of 0.64 (section 3.1). Despite including no physical or mechanistic processes, MLR obtains a realistic PNV distribution closer to BIOME 6000 data. A $\kappa$ coefficient computed between RF99 and BIOME 6000 is not possible given the different biomes vegetation types. The RF99 map appears more heterogeneous than MLR. Some of these local details are disputable and correspond more to current observed vegetation rather than potential vegetation. For example, RF99 sees the Landes forests (in the region of Landes in France) which have been mainly planted by humans. MLR sees also a dominance of forests, but with large uncertainty.

Indeed, MLR does not only provide a vegetation distribution, since we obtain an occurrence probability by megabiome. Occurrence probabilities allows us to estimate the uncertainty of the modelled PNV distribution, taking into account the megabiome with second highest occurrence probability. This second dominant megabiome often appears in agreement in regions where the first dominant megabiome is different from other databases. The percentage of agreement with BIOME 6000 increases from 69.5% (see section 3.1) to 89.9%, taking into account the second dominant megabiome. In agreement with RF99 or BIOME4, the map of the second dominant megabiome modelled by MLR (not shown) shows tundra and boreal forests in northeastern Europe (section 3.3), tropical forest in equatorial region and cold desert in Andes capturing the effect of local-scale topography through the high-resolution CRU climatologies. An uncertainty index UI can also be computed from the difference between the two highest occurrence probabilities (figure 3):

$$\mathrm{UI} = \frac{1-(p_1-p_2)}{p_1} \qquad (2)$$

where $p_x$ is the occurrence probability and $x$ the rank of the probability ranging from 1 (the highest probability) to $m$ (the lowest probability) with $m$ the number of megabiomes.

A first dominant megabiome with a probability close to the probability of the second dominant megabiome has an uncertainty close to 1 and vice versa. This index appears very useful in bringing some confidence in the PNV modelled by MLR and pointing out the limits of our method:

- This uncertainty index allows us to target the regions needing some data to improve the PNV distribution, such as in South America. Indeed, the main limit of the MLR method lies in the training data (BIOME 6000 in this case). The modelled PNV by MLR highly depends on the abundance and geographical distribution of data points. If a megabiome is absent or over/under-represented, this will have a significant impact on the modelled PNV by MLR. Nevertheless, a calibration of MLR over the globe provides a geographical robustness to the statistical model. The PNV predicted by MLR in regions with no or less BIOME 6000 data appears consistent with climatic patterns (e.g., MLR shows similarities with modern biome reconstructions from Marchant *et al* 2009 in several regions of South America).
- This index highlights the regions where the modelled vegetation is to be taken with caution (as in western US) because the climatic signal alone is not sufficient

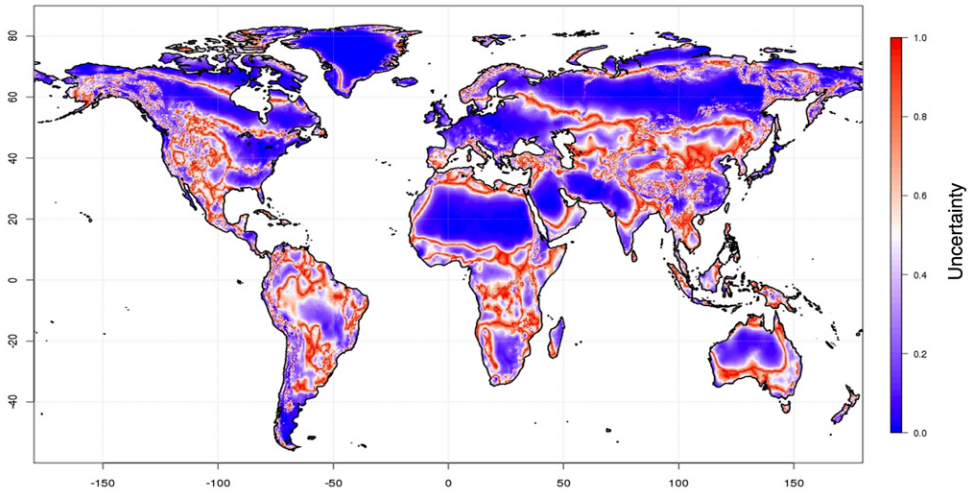

**Figure 3.** The uncertainty index of the maximum occurrence probability predicted by MLR: 1 means 'high uncertainty' and 0 means 'no uncertainty'.

to distinguish a dominant megabiome. The role of fires, herbivory or soil properties has been clearly identified for vegetation (e.g., Sankaran *et al* 2005 for savannas) and could be included as predictors in our statistical model.

Like any database, the disadvantages of our statistical approach should be discussed to better constrain its application. The 'vegetation–climate' relationship estimated by MLR from BIOME 6000 modern data is implicitly constrained by an atmospheric $CO_2$ concentration of about 360 ppm. As a prospect, exporting this relationship in different climatic conditions leads to a distribution ignoring the crucial effect of $CO_2$ on vegetation (Harrison and Prentice 2003, Woillez *et al* 2011). Moreover, MLR do not simulate soil–vegetation–atmosphere interactions such as photosynthesis, growth and competitiveness of plants, which may be more constant at the biome level. Vegetation models allow us to provide characteristics of vegetation as leaf area index (LAI) or net primary productivity (NPP). Statistical modelling of vegetation is an interesting and complementary alternative to process-based vegetation models.

Finally, the PNV modelled by MLR cannot claim to be fully independent of human influences. MLR is mainly based on climatological data between 1961 and 1990, impacted by human activities through climate change. Moreover, BIOME 6000 data includes modern data referring to samples dated within the past thousand years (most of pollen samples falls within the past 500 years (Bigelow *et al* 2003)). Man has intensively used lands for thousands of years (for example, in ancient Greece, during the Western agricultural revolution in the Middle Ages or more recently with the Green Revolution between 1960 and 1980). To warrant the 'potential' feature of the modelled vegetation by MLR, it could be relevant to calibrate MLR on BIOME 6000 data from the Holocene (−6 ka). At this period, the land-use was limited to a few scattered subtropical farm households (e.g., in China or South America).

For details about these last remarks, the method and the used data, we refer the reader to Levavasseur *et al* (2012). All final data (megabiomes and occurrence probabilities) are in an attached supplementary NetCDF file (available at stacks.iop.org/ERL/7/044019/mmedia).

To conclude, for the modern period, BIOME 6000 can be confidently considered as reference data collected in areas with less possible human activity, although ensuring data not impacted by humans is difficult. Accounting all our observations and statistical indices, MLR models the most realistic PNV on the regions covered by BIOME 6000. Over the rest of the world, MLR models a vegetation distribution consistent with climatic signal. The MLR method is a fast and robust alternative in vegetation modelling with several advantages. The modelled PNV map is (i) directly and only based on vegetation (BIOME 6000) and climatological (CRU) data; (ii) not subjective and independent of any vegetation model; (iii) easily updatable as soon as additional data is made available.

## Acknowledgments

We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://eca.knmi.nl). G Levavasseur is supported by UVSQ, D M Roche and M Vrac by INSU/CNRS.

## Appendix A. Growing degree day at 5 °C downscaling

As GDD5 is built from temperatures, we choose to apply the statistical downscaling method developed by Vrac *et al* (2007) for temperature downscaling and based on the use of a Generalized Additive Model (GAM, Hastie and Tibshirani 1990). GAM models statistical relationships between local-scale observations over Europe: the high-resolution GDD5 climatology built from the ECA&D times-series (Haylock *et al* 2008); and global variables: the topography from ETOPO2 (Amante and Eakins 2009), the temperature from CRU (New *et al* 2002), and the GDD5 built from NCEP/NCAR (see section 2.2.1). GAM represents the expectation of an explained variable $Y$ (the predictand, i.e., the GDD5 from ECA&D in our case) by a sum of nonlinear functions $f_k$, conditionally on explanatory variables (the predictors, i.e., the topography, the temperature and the GDD5 from NCEP/NCAR) $X_k$ (Hastie and Tibshirani 1990):

$$E(Y_i|X_{k,k=1\cdots n}) = \sum_{k=1}^{n} f_k(X_{i,k}) + \epsilon, \qquad \text{(A.1)}$$

where $\epsilon$ is the residual or error, $n$ is the number of predictors and $i$ is the grid-cell. To use GAM, we need to define the distribution family of the explained variable. For simplicity, Vrac *et al* (2007) assumed that temperature has a Gaussian distribution so we assume that GDD5 too, which implies a zero-mean Gaussian error $\epsilon$ (Hastie and Tibshirani 1990). Then, we define the nonlinear functions as cubic regression splines (piecewise by third-degree polynomials). Finally, any SDM needs a calibration/projection procedure. The calibration is the fitting process of the splines over Europe in our case. Afterwards, we project over the world to predict a high-resolution global GDD5 climatology.

Instead of a simple bilinear interpolation of the NCEP/NCAR GDD5, we use GAM to geographically extrapolate the characteristics of the ECA&D GDD5 over Europe to the world. For more details we refer the reader to Vrac *et al* (2007) and Martin *et al* (2012). We perform this analysis within the statistical programming environment R (R Development Core Team 2011) and its ‘mgcv’ package (Wood 2006).

## Appendix B. Statistical indices used for model selection

*The Bayesian Information Criterion (BIC).* The BIC (equation (B.1)) is a particular form of the Akaike Information Criterion (Sakamoto *et al* 1986), developed by Schwarz (1978) and defined by:

$$\text{BIC} = -2 \times \text{LL} + P \times \log(n) \qquad \text{(B.1)}$$

where $n$ corresponds to the number of BIOME 6000 data points ($n = 6091$), $P$ is the number of parameters in the fitted model ($P = n \times (m - 1)$) and LL is the log-likelihood of the fitted model. This criterion measures the goodness-of-fit between the statistical model and the data, balancing the risk of over-fitting. The BIC includes a penalty term depending on the sample size ($n$) and on the dimension of the model ($P$). The smaller the BIC, the better the model.

*Pseudo-$R^2$.* The $R^2$ is a classical statistical index in ordinary least squares regression that is often used as a goodness-of-fit measure. In logistic regression, an equivalent statistic to $R^2$ does not exist. However, to evaluate the goodness-of-fit of logistic models, several ‘pseudo-$R^2$’ (ranging from 0 to 1) have been proposed. Among the different approaches, the McFadden’s pseudo-$R^2$ is often used for its simplicity of calculation (equation (B.2)) and interpretation. It is defined by (Menard 2000, Shtatland *et al* 2002):

$$R^2 = 1 - \frac{\text{LL}}{\text{LL}_{\text{null}}} \qquad \text{(B.2)}$$

where LL is the log-likelihood of the selected model (i.e., with selected predictors) and $\text{LL}_{null}$ the log-likelihood for the null-model (i.e., with intercept only). The ratio of log-likelihoods suggests the level of improvement over the null-model offered by the involved predictors. A small ratio of likelihoods indicates that the full model is far better than the null-model. In terms of pseudo-$R^2$, the closer the $R^2$ is to 1, the better the agreement with data is.

*The kappa statistic.* The $\kappa$ coefficient measures the quality of the agreement (Cohen 1960, Fleiss *et al* 1969) between the modelled PNV by MLR in each BIOME 6000 location (figure 1(b)) and the BIOME 6000 data (figure 1(a)). This index can take values between 0 and 1 and is based on a simple counting of matching and non-matching points in a matrix used to represent errors in assigning classes (see appendix A of Levavasseur *et al* (2011)). The closer the $\kappa$ coefficient is to 1, the better the agreement with data is. The kappa statistic is often used for spatial comparison of categorical variables, such as vegetation (Monserud and Leemans 1992).

*The Brier score.* The Brier score was developed by Brier 1950 to assess the accuracy of probabilistic forecasts. As MLR provides probabilities of occurrence of different megabiomes, this score is well adapted here. It measures the average squared deviation between predicted probabilities for a set of events and their binary outcomes (0 if the event does not happen and 1 if it happens). For a multinomial variable, the Brier score is defined by:

$$\text{BS} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} (p_{i,j} - o_{i,j})^2 \in [0;\ m] \qquad \text{(B.3)}$$

where $n$ is the number of BIOME 6000 data points and $m$ is the number of megabiomes. $p_{i,j}$ corresponds to the predict probability of the $j$th megabiomes at the $i$th point/location and $o_{i,j}$ is the corresponding binary outcome for this point. The Brier score can take values between 0 and $m$. A lower score represents higher accuracy of the prediction. The Brier score can also be reduced in two other ways:

- Taking into account all $m$ megabiomes by grid-cell/location, we obtain a map of Brier scores:

$$\text{BS}(i) = \sum_{j=1}^{m} (p_j - o_j)^2 \in [0;\ m]. \qquad \text{(B.4)}$$

- Taking into account all $n$ grid-cells/locations by megabiome, we obtain $m$ Brier scores (i.e., for each $j$ megabiome):

$$\text{BS}(j) = \frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2 \in [0;\ 1]. \qquad \text{(B.5)}$$

## References

Amante C and Eakins B 2009 ETOPO1 1 arc-minute global relief model: procedures, data sources and analysis *NOAA Technical Memorandum NESDIS* NGDC-24 (Boulder, CO: NGDC)
Bigelow N H *et al* 2003 *J. Geophys. Res.* **108** 8170–95
Brier G 1950 *Mon. Weather Rev.* **78** 1–3
Cha G 1997 *J. For. Res.* **2** 147–52
Cohen J 1960 *Educ. Psychol. Meas.* **20** 37–46
Crucifix M, Betts R and Hewitt C 2005 *Glob. Planet. Change* **45** 295–312
Fleiss J, Cohen J and Everitt B 1969 *Psychol. Bull.* **72** 323–7
Guiot J 2012 personal communication
Harrison S and Prentice C 2003 *Glob. Change Biol.* **9** 983–1004
Harrison S, Yu G, Takahara H and Prentice I 2001 *Nature* **413** 129–30
Hastie T and Tibshirani R 1990 *Generalized Additive Models* 1st edn (London: Chapman and Hall, CRC Press)
Haxeltine A and Prentice I 1996 *Glob. Biogeochem. Cycles* **10** 693–709
Haylock M, Hofstra N, Klein Tank A, Klok E, Jones P and New M 2008 *J. Geophys. Res.* **113** D20119
Hilbe J 2009 *Logistic Regression Models* 1st edn (London: Chapman and Hall, CRC Press)
Hosmer D and Lemeshow S 2000 *Applied Logistic Regression* 2nd edn (New York: Wiley)
Kalnay E *et al* 1996 *Bull. Am. Meteorol. Soc.* **77** 437–71
Kaplan J, Krumhardt K and Zimmermann N 2009 *Quat. Sci. Rev.* **28** 3016–34
Kaplan J O *et al* 2003 *J. Geophys. Res.* **108** 8171–88
Levavasseur G, Roche D, Vrac M, Paillard D and Guiot J 2012 *Glob. Planet. Change* in review
Levavasseur G, Vrac M, Roche D M, Paillard D, Martin A and Vandenberghe J 2011 *Clim. Past* **7** 1647–92
Loveland T and Belward A 1997 *Int. J. Remote Sens.* **18** 3289–95
Loveland T, Reed B, Brown J, Ohlen D, Zhu Z, Yang L and Merchant J 2000 *Int. J. Remote Sens.* **21** 1303–30
Lüthi D *et al* 2008 *Nature* **453** 379–82
Marchant R *et al* 2009 *Clim. Past* **5** 369–461
Martin A, Vrac M, Paillard D, Dumas C and Kageyama M 2012 *Clim. Dyn.* in review
Menard S 2000 *Am. Stat.* **54** 17–24
Monserud R and Leemans R 1992 *Ecol. Modelling* **62** 275–93
New M, Lister D, Hulme M and Makin I 2002 *Clim. Res.* **21** 1–25
Ni J, Harrison S, Prentice I, Kutzbach J and Sitch S 2006 *Ecol. Modelling* **191** 469–86
Notaro M 2008 *J. Clim.* **30** 845–54
Notaro M, Zhengyu L, Gallimore R, Vavrus S, Kutzbach J, Prentice I and Jacob R 2005 *J. Clim.* **18** 3650–71
Pickett E *et al* 2004 *J. Biogeogr.* **31** 1381–444
Prentice C, Guiot J, Huntley B, Jolly D and Cheddadi R 1996 *Clim. Dyn.* **12** 185–94
Prentice I, Cramer W, Harrison S, Leemans R, Monserud R and Solomon A 1992 *J. Biogeogr.* **19** 117–34
Prentice I and Jolly D 2000 *J. Biogeogr.* **27** 507–19
Ramankutty N and Foley J 1999 *Glob. Biogeochem. Cycles* **13** 997–1027
R Development Core Team 2011 *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing) (www.R-project.org)
Sakamoto Y, Ishiguro M and Kitagawa G 1986 *Akaike Information Criterion Statistics* illustrated edn (Norwell, MA: Kluwer Academic Publishers)
Sankaran M *et al* 2005 *Nature* **438** 846–9
Schwarz G 1978 *Ann. Stat.* **6** 461–4
Shtatland E, Kleinman K and Cain E 2002 One more time about $R^2$ measures of fit in logistic regression *Proc. 15th NESUG Statistics, Data Analysis and Econometrics* pp 222–6
Sitch S *et al* 2003 *Glob. Change Biol.* **9** 161–85
Tang G, Shafer S, Bartlein P and Holman J 2009 *Ecol. Modelling* **220** 1481–91
Vrac M, Marbaix P, Paillard D and Naveau P 2007 *Clim. Past* **3** 669–82
Woillez M N, Kageyama M, Krinner G, de Noblet-Ducoudré N, Viovy N and Mancip M 2011 *Clim. Past* **7** 557–77
Wood S 2006 *Generalized Additive Models: An Introduction with R* 1st edn (London: Chapman and Hall, CRC Press)
Yee T 2010a *J. Stat. Softw.* **32** 1–34
Yee T 2010b *R Package Version 0.8-1* (http://CRAN.R-project.org/package=VGAM)
Yee T and Wild C 1996 *J. R. Stat. Soc.* **58** 481–93