Effects of experimental protocol on global vegetation model accuracy: A comparison of simulated and observed vegetation patterns for Asia

Guoping Tang\textsuperscript{a,}\textsuperscript{*}, Sarah L. Shafer\textsuperscript{b}, Patrick J. Bartlein\textsuperscript{a}, Justin O. Holman\textsuperscript{c}

\textsuperscript{a} Department of Geography, University of Oregon, Eugene, OR 97403, USA
\textsuperscript{b} U.S. Geological Survey, Corvallis, OR 97331, USA
\textsuperscript{c} TerraSeer, Inc., Ann Arbor, MI 48104, USA

\textbf{Article info}

\textbf{Article history:}
Received 18 July 2008
Received in revised form 16 March 2009
Accepted 18 March 2009

\textbf{Keywords:}
Climatology
Model validation
Map comparison
Accuracy assessment
BIOME4
Vegetation
Asia

\textbf{Abstract}

Prognostic vegetation models have been widely used to study the interactions between environmental change and biological systems. This study examines the sensitivity of vegetation model simulations to: (i) the selection of input climatologies representing different time periods and their associated atmospheric CO\textsubscript{2} concentrations, (ii) the choice of observed vegetation data for evaluating the model results, and (iii) the methods used to compare simulated and observed vegetation. We use vegetation simulated for Asia by the equilibrium vegetation model BIOME4 as a typical example of vegetation model output. BIOME4 was run using 19 different climatologies and their associated atmospheric CO\textsubscript{2} concentrations. The Kappa statistic, Fuzzy Kappa statistic and a newly developed map-comparison method, the Nomad index, were used to quantify the agreement between the biomes simulated under each scenario and the observed vegetation from three different global land- and tree-cover data sets: the global Potential Natural Vegetation data set (PNV), the Global Land Cover Characteristics data set (GLCC), and the Global Land Cover Facility data set (GLCF). The results indicate that the 30-year mean climatology (and its associated atmospheric CO\textsubscript{2} concentration) for the time period immediately preceding the collection date of the observed vegetation data produce the most accurate vegetation simulations when compared with all three observed vegetation data sets. The study also indicates that the BIOME4-simulated vegetation for Asia more closely matches the PNV data than the other two observed vegetation data sets. Given the same observed data, the accuracy assessments of the BIOME4 simulations made using the Kappa, Fuzzy Kappa and Nomad index map-comparison methods agree well when the compared vegetation types consist of a large number of spatially continuous grid cells. The results of this analysis can assist model users in designing experimental protocols for simulating vegetation.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Over recent decades, many studies have examined the impacts of climate change on terrestrial vegetation dynamics (\textit{e.g.} Prentice et al., 1993; Harrison and Prentice, 2003) and the global carbon cycle (\textit{e.g.} White et al., 2000; Woodward and Lomas, 2004). To assist with these studies, a number of prognostic vegetation models have been developed and used to examine the interactions between environmental change and biological systems (Bolliger et al., 2000). Vegetation models such as IBIS (Foley et al., 1996), BIOME4 (Kaplan et al., 2003), LPJ (Sitch et al., 2003), CENTURY (Parton et al., 1987) and BIOME-BGC (Running and Coughlan, 1988), have greatly improved our ability to understand the response of terrestrial vegetation to past and future environmental variation at global-to-regional scales. However, evaluations of the performance of model-based simulations are subject to a number of limitations, both inherent in the models themselves and in the input data used to run the models (Allen et al., 2001; Barrett et al., 2001). For example, many equilibrium vegetation models assume that climate is the primary factor controlling the distribution of terrestrial vegetation, and so model-simulated vegetation is consequently highly dependent on the quality and characteristics of the specific input (or baseline) climatology data, including the number of years of climate data included in the climatology (Kickert et al., 1999).

Remotely sensed global land- and tree-cover data sets, such as the Advanced Very High Resolution Radiometer (AVHRR) Pathfinder Land data, are frequently used to evaluate the ability of vegetation models to simulate global terrestrial vegetation because these data map the pattern of vegetation over large areas (Turner et al., 1993; Gould, 2000). However, errors inherent in remotely sensed image processing steps, such as image classification and geo-registration, may misclassify the vegetation pattern...
in a given area (Campbell, 2002). In addition, remotely sensed image classifications developed for different purposes may put different emphases on the categories and characteristics of the post-classified data. For example, NOAA’s 1992–93 AVHRR data have been used to produce both 1-km global land cover characteristics data with 97 different land cover types (Loveland et al., 2000) and continuous field tree-cover data with only three tree-cover classes (DeFries et al., 2000; Hansen et al., 2000). The accuracy assessment of simulated vegetation may vary significantly depending on the choice of the observed vegetation data set used to evaluate the simulated vegetation.

A number of quantitative map-comparison techniques have been used to assess the ability of vegetation models to simulate observed vegetation. These techniques employ such statistics as the Kappa statistic (Cohen, 1960; Congalton and Green, 1999), Tau (Ma and Redmond, 1994), Kappa-for-location (Pontius, 2000), and Fuzzy Kappa (Hagen, 2003), and are used in raster-based categorical data comparison through pixel-to-pixel arithmetic. Each approach has its own strengths and weaknesses for evaluating a model’s performance partially because model results are often spatiotemporally autocorrelated (Tang, 2008; Tang and Bartlein, 2008). For example, the Kappa statistic may greatly underestimate the similarity of two maps if they display a similar data pattern but that pattern is slightly offset from one map to the other (Foody, 1992, 2002). The assessed accuracy of simulated vegetation will vary depending on the choice of which quantitative techniques are used to compare maps of simulated and observed vegetation.

The goal of this study is to examine the effects of three elements of experimental protocols on the evaluation of global vegetation model accuracy: (1) the choice of the input climatology (and its associated atmospheric CO2 concentration), (2) validation data selection, and (3) the map-comparison method. We use vegetation simulated with the equilibrium model BIOME4 (Kaplan et al., 2002, 2003) as a typical example of the output from a vegetation model. The geographic focus of this study is most of Asia, ranging from 60.0°E to 150.0°E and from 8.0°N to 80.0°N. This region was chosen because it contains a diversity of terrestrial vegetation and climate zones, and few detailed studies assessing vegetation simulations have been done for this area (e.g. Song et al., 2005). Methodologically, the well-known Kappa statistic, Fuzzy Kappa statistic (hereafter Fuzzy Kappa) and a newly developed method for map comparison, the Nomad index (Holman, 2004), are employed to quantify the agreement between BIOME4-simulated biomes and those derived from three global land- and tree-cover data sets. Throughout this analysis, we hope to provide model users with guidance in designing experimental protocols for vegetation model simulations.

2. Materials and methods

2.1. The BIOME4 model and baseline climatology data sets

BIOME4 (version 2b1) is an equilibrium, coupled biogeography and biogeochemistry vegetation model that simulates global vegetation in the form of 13 plant functional types (PFTs) that are combined to form 27 biomes (Table 1; Kaplan et al., 2002, 2003). The model also calculates a variety of other variables, such as net primary productivity and leaf area index for each PFT under a prescribed global atmospheric CO2 concentration. BIOME4 has been employed in a number of studies of past, present and potential future vegetation patterns (e.g. Bigelow et al., 2003; Diffenbaugh et al., 2003; Song et al., 2005).

The input variables required to run BIOME4 include monthly mean temperature (°C), monthly mean total precipitation (mm), monthly mean sunshine (%), soil water holding capacity (mm) for each of two soil layers, a conductivity index of water movement through the soil column (mm/day), and atmospheric CO2 concentration (ppm). We used the CRU TS 2.0 data set from the Climatic Research Unit (CRU), University of East Anglia (U.K.), to create the input climatologies for running BIOME4. The CRU TS 2.0 data set is supplied on a 0.5-degree global land grid at a monthly time-step for 1901–2000 and builds upon several previous CRU gridded data sets (New et al., 1999, 2000; Mitchell et al., 2004). Other global climate data sets of long-term averages or climatologies are available for equilibrium-model simulations (e.g. WorldClim, Hijmans et al., 2005) but for this study we required monthly time-series data. Soil data were obtained from the derived soil properties defined in the Food and Agriculture Organization (FAO) digital soil map of the world (FAO, 1995). Annual atmospheric CO2 values for 1901–1995 were provided by the Carbon Cycle Model Linkage Project (Kicklighter et al., 1999).

We designed 19 scenarios to analyze the sensitivity of BIOME4 simulations to different climatologies and their associated atmospheric CO2 concentrations. The 19 monthly mean climatologies used to run BIOME4 are derived from the CRU TS 2.0 data set and each of them extends backward in time from December of 1992 for 2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85 and 90 years respectively. The year 1992 was chosen as the common end point for the climatologies because all three observed vegetation data sets used in this study were developed from remotely sensed images taken over the period of April 1992 to March 1993. The associated atmospheric CO2 concentrations for each BIOME4 simulation were developed by averaging annual CO2 values for each of the 19 different climatology time periods. We also designed 8 scenarios to examine the sensitivity of BIOME4 simulations to different 30-year monthly mean climatologies (1963–1992, 1961–1990, 1951–1980, 1941–1970, 1931–1960, 1921–1950, 1911–1940 and 1901–1930) and their associated atmospheric CO2 concentrations following the same approach.

2.2. Global land cover and forest cover data

We used one tree-cover and two global land cover data sets to investigate how the selection of observed vegetation data influences the accuracy assessment of the model-based simulations. The data sets include (i) the global potential natural vegetation (PNV) data (Ranamukutty and Foley, 1999), (ii) the 1-km global land cover characteristics (GLCC) data (version 2.0) (Loveland et al., 2000), and (iii) the global AVHRR continuous field tree-cover (GLCF) data (DeFries et al., 2000; Hansen et al., 2000). Each observed vegetation data set was compared to each of the 19 BIOME4 simulations.

The gridded global PNV data describe the distribution of global potential natural vegetation patterns that would most likely exist in the absence of human activities (Ranamukutty and Foley, 1999).
For regions not dominated by human land use, the PNV vegetation types were classified into 15 biome types (Table 2) at a spatial resolution of 0.5°, which matches the resolution at which biomes were simulated by BIOME4 for this study. The GLCC data were generated from IGBP 1-km AVHRR 10-day composites from April 1992 through March 1993 (Eidenshink and Faundeen, 1994), and the data were classified into 97 global land cover types at a 30 arc-second spatial resolution. The GLCF data were obtained by combining the "Global maps of proportional cover for three vegetation characteristics" data (DeFries et al., 1998) with the "Global land cover classification" data (Hansen et al., 2000). Both of these data sets were derived at a 1-km spatial resolution from NOAA's AVHRR data acquired in 1992–1993. The GLCC data used in this study are at a 30 arc-second spatial resolution and consist of continuous fields of vegetation characteristics divided into three classes: (i) percent tree cover with values ranging from 10 to 80 percent; (ii) percent tree cover less than 10 percent; and (iii) non-vegetated areas (coded as "254") (DeFries et al., 2000; Hansen et al., 2000).

### 2.3. Data processing

Unlike the global PNV data, which have the same spatial resolution (0.5°) as the CRU TS 2.0 data set, the global GLCC and GLCF data have a much finer spatial resolution. In order to compare the BIOME4-simulated biome types with the land cover classes in the GLCC data and the tree-cover classes in the GLCF data, we first reclassified the 30 arc-second GLCC and GLCF data into 0.5° gridded data using a nearest-neighbor algorithm. For each 0.5° grid point we assigned the value of its most-likely neighbor (i.e. the mode of land cover types) as the BIOME4 biome types, such as "heath scrub" (which consists of 23 grid cells in the GLCC data), are often clustered in specific areas and represented by relatively few grid cells. Thus, they do not have a large effect on the accuracy assessments of the BIOME4 simulations. In addition, given that BIOME4 simulates potential vegetation in the absence of human activity, we excluded from our analysis all grid cells in which the GLCC land cover types represent human activities (e.g. crops, urban).

There is good agreement between the BIOME4 biome categories and the PNV categories (Tables 1 and 2). We classified vegetation types from these two data sets into 6 biome types: forest cover, tropical/temperate savanna, grassland, tropical/temperate shrubland, tundra, and barren/desert (Table 4). In order to compare the BIOME4-simulated biomes with those classified in the GLCF data, land simulated under the 1961–1990 30-year climatology consists of only three grid cells and was combined with tropical savanna.

### Table 2

Original biome types classified in the PNV data. (Ramankutty and Foley, 1999).

<table>
<thead>
<tr>
<th>Biome types classified in the PNV data</th>
<th>BIOME4 biome codes</th>
<th>GLCC land cover codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Tropical evergreen forest/woodland</td>
<td>9) Savanna</td>
<td>9) Savanna</td>
</tr>
<tr>
<td>2) Tropical deciduous forest/woodland</td>
<td>10) Grassland/steppe</td>
<td>10) Grassland/steppe</td>
</tr>
<tr>
<td>3) Temperate broadleaf evergreen forest/woodland</td>
<td>11) Dense shrubland</td>
<td>11) Dense shrubland</td>
</tr>
<tr>
<td>4) Temperate needleleaf evergreen forest/woodland</td>
<td>12) Open shrubland</td>
<td>12) Open shrubland</td>
</tr>
<tr>
<td>5) Temperate deciduous forest/woodland</td>
<td>13) Tundra</td>
<td>13) Tundra</td>
</tr>
<tr>
<td>6) Boreal evergreen forest/woodland</td>
<td>14) Desert</td>
<td>14) Desert</td>
</tr>
<tr>
<td>7) Boreal deciduous forest/woodland</td>
<td>15) Polar desert/rock/ice</td>
<td>15) Polar desert/rock/ice</td>
</tr>
</tbody>
</table>

### Table 3

Reclassified BIOME4 biome types and GLCC land cover types.

<table>
<thead>
<tr>
<th>Reclassified biome types</th>
<th>BIOME4 biome codes</th>
<th>GLCC land cover codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Tropical evergreen forest</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>9, 10, 11, 12, 13, 14, 15</td>
</tr>
<tr>
<td>2) Tropical deciduous forest</td>
<td>2, 3, 4, 5</td>
<td>6, 7, 8, 9, 10, 11</td>
</tr>
<tr>
<td>3) Temperate deciduous/forest</td>
<td>4, 5</td>
<td>6, 7, 8</td>
</tr>
<tr>
<td>4) Warm mixed forest</td>
<td>6</td>
<td>7, 8</td>
</tr>
<tr>
<td>5) Cool mixed forest</td>
<td>7, 8</td>
<td>11, 12</td>
</tr>
<tr>
<td>6) Cold conifer forest</td>
<td>8</td>
<td>13, 14</td>
</tr>
<tr>
<td>7) Evergreen taiga/montane forest</td>
<td>10</td>
<td>15, 16, 17, 18</td>
</tr>
<tr>
<td>8) Deciduous taiga/montane forest</td>
<td>11</td>
<td>19, 20, 21, 22</td>
</tr>
<tr>
<td>9) Tropical/savanna and grassland</td>
<td>12, 19</td>
<td>23, 24, 25, 26</td>
</tr>
<tr>
<td>10) Tropical/temperate shrubland</td>
<td>13, 14</td>
<td>27, 28</td>
</tr>
<tr>
<td>11) Temperate/temperate shrubland</td>
<td>15</td>
<td>29, 30</td>
</tr>
<tr>
<td>12) Temperate savanna and grassland</td>
<td>16, 17, 18</td>
<td>31, 32</td>
</tr>
<tr>
<td>13) Barren and desert</td>
<td>19</td>
<td>33, 34</td>
</tr>
<tr>
<td>14) Steppe tundra</td>
<td>20</td>
<td>35, 36</td>
</tr>
<tr>
<td>15) Shrub tundra</td>
<td>21</td>
<td>37, 38</td>
</tr>
<tr>
<td>16) Dwarf and prostrate shrub and moss tundra</td>
<td>22</td>
<td>39, 40</td>
</tr>
<tr>
<td>17) Grassland</td>
<td>23</td>
<td>41, 42</td>
</tr>
<tr>
<td>18) Mixed forest</td>
<td>24</td>
<td>43, 44</td>
</tr>
</tbody>
</table>

### Table 4

Reclassified biome types and PNV types for comparison.

<table>
<thead>
<tr>
<th>Reclassified biome types</th>
<th>BIOME4 biome codes</th>
<th>PNV biome codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Forest cover</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>9, 10, 11</td>
</tr>
<tr>
<td>2) Tropical/temperate savanna</td>
<td>12, 16</td>
<td>9</td>
</tr>
<tr>
<td>3) Grassland</td>
<td>19, 20</td>
<td>10</td>
</tr>
<tr>
<td>4) Tropical/temperate shrubland</td>
<td>13, 14</td>
<td>11, 12</td>
</tr>
<tr>
<td>5) Tundra</td>
<td>22, 23, 24, 25, 26</td>
<td>13</td>
</tr>
<tr>
<td>6) Barren and desert</td>
<td>21, 27, 28</td>
<td>14, 15</td>
</tr>
</tbody>
</table>
we grouped the BIOME4-simulated biomes into three categories: forest biomes, grassland and savanna biomes, and desert biomes including barren areas (Table 5). As with the comparison of BIOME4 vs. GLCC, the comparison of BIOME4 vs. GLCF data also excluded all grid cells where the GLCC land cover types reflected human activities.

2.4. Map-comparison methods

2.4.1. The Kappa statistic

The Kappa statistic (Cohen, 1960) has gained widespread use in assessing model-simulated vegetation distribution (e.g. Prentice et al., 1992; Diffenbaugh et al., 2003). The advantage of the Kappa statistic is that it takes into account chance agreement, regardless of the number of categories in the maps being compared. In addition, it is easy to calculate and is an intuitive measure of agreement (Monserud and Lemmings, 1992; Foody, 2002). For each pair of compared observed and simulated vegetation data sets, an error matrix for all observed and simulated vegetation combinations, is constructed. Then, for each category in the constructed error matrix, the Kappa statistic is calculated by the following equation:

\[
k_i = \frac{\left( p_{irow} - p_{icol} \right)}{\left( \left( p_{irow} + p_{icol} \right)/2 - p_{icol} \right)} (1)
\]

where \( p_{irow} \) is the row total for each category \( i \); \( p_{icol} \) is the column total for each category \( i \); and \( p_{i} \) is the individual entry for the row and column on the main diagonal of constructed error matrix. The overall agreement between two compared maps is estimated by the formula:

\[
k = \frac{p_0 - p_e}{1 - p_e} (2)
\]

where \( p_0 = \sum_{i=1}^{c} p_{i} \); \( p_e = \sum_{i=1}^{c} p_{irow} p_{icol} \); and \( c \) is the number of categories in each data set.

2.4.2. The Fuzzy Kappa statistic

Like the Kappa statistic, the Fuzzy Kappa statistic is also a cell-by-cell-based map-comparison approach. However, in order to distinguish minor differences from major differences, the Fuzzy Kappa assesses two types of fuzziness: fuzziness of category and fuzziness of location. Category fuzziness refers to the ordinal similarity among all categories on a map. Location fuzziness refers to the fact that the spatial location of a category on a map is not always precise (Hagen, 2003; Hagen-Zanker et al., 2005). For each cell, a local measure of similarity can be calculated based on a two-way comparison. An overall Fuzzy Kappa statistic can be obtained by averaging the similarity calculated for all grid cells, which yields a result between 0 (for total disagreement) and 1 (for identical maps). The formula for Fuzzy Kappa is identical in form to Eq. (2) but different in the calculation of the expected similarity (\( P_e \)), which is expressed by the formula:

\[
P_e = \sum_{i=1}^{R} E(i) \times M(d_i) (3)
\]

where \( R \) is the number of the furthest ring (cells that are at the same distance from a central cell are said to form a neighborhood ring); \( M \) is the fuzzy membership function that describes the similarity of a category in two compared maps; \( d_i \) is the radius of ring \( i \); and \( E(i) \) refers to the probability of matching central cells and is calculated separately according to the Kappa statistic. In this study, we calculate location fuzziness using an exponential decay membership function with a halving distance of 0.5 and a neighborhood radius of 1 grid cell. Category fuzziness is not considered because it produces measured agreement that is very similar to the agreement produced using location fuzziness. A Fuzzy Kappa statistic for all individual categories in two maps can be calculated through the creation of “temporary category similarity matrices,” in which all categories are set equal to each other except for the category being considered (Hagen, 2003).

2.4.3. The Normalized Minimum Agreement Distance (nomad) index

The Nomad index (Holman, 2004) was developed to quantify the agreement of spatial patterns between maps and to overcome map coregistration issues (e.g. Costanza, 1989). For each vegetation category in the two map raster data sets being compared, the calculation of the Nomad index follows three steps: (i) determine a set of grid-cell pairs (one grid cell in map A, one grid cell in map B) that minimizes the total geographic distance between all possible grid-cell pairs; (ii) calculate the average geographic distance for this minimum-distance set of pairs and the overall average grid-cell-to-grid-cell distance for the category; and (iii) compare the average minimum geographic distance (determined in i) with the overall average category-to-grid-cell distance for the category to quantify the level of map agreement (Holman, 2004). If the minimum distance between pairs for a category is small compared to the overall average distance of a category, then map agreement will relatively high and vice versa. For each category, the values of the Nomad index range from near or below 0 for poor agreement to 1 for perfect agreement. The overall map agreement is determined by calculating a frequency-weighted average of individual Nomad values for each vegetation category.

The individual Nomad index \( n(c) \) for category \( c \) in two maps is calculated according to the following equation:

\[
n(c) = 1 - \left( \frac{1}{d_{avg}} \right) (4)
\]

where \( d(c) = \min \left( \sum d_{ab} \right) \), which stands for the minimum geographic distance for all possible grid-cell comparisons \( d_{ab} \); \( d_{avg} \) is the distance from grid cell \( a \) in map A to its match \( b \) in map B; \( d_{avg} \) is the overall average distance of \( d_{ab} \); and \( c \) is the number of grid cells in category \( c \), where the distances are determined using an optimization algorithm (Holman, 2004). The Nomad index can be intuitively thought of as a measure of how much the pixels on one map would have to be moved to match the pixels on the other map. The overall Nomad index \( N \) for two maps with \( k \) categories is expressed as the following equation:

\[
N = \frac{\sum n(c)}{k} (5)
\]

2.4.4. The rating system used in the accuracy assessment

For the Kappa statistic and Fuzzy Kappa, we used the accuracy rating system of Landis and Koch (1977) and Monserud and Lemmings (1992) where values greater than 0.75 indicate very good-to-excellent agreement, values between 0.40 and 0.75 indicate fair-to-good agreement, and values of 0.40 or less indicate poor agreement. Values close to 0.0 suggest that the agreement is no better than would be expected by chance. For the Nomad index, Holman (2004) suggests values greater than 0.75 indicate very
good-to-excellent agreement, values between 0.60 and 0.75 indicate fair-to-good agreement, and values less than 0.60 indicate relatively poor agreement.

3. Results and discussion

3.1. Effect of different input climatologies and their associated CO₂ concentrations on simulated vegetation

A number of patterns emerge when the BIOME4 vegetation simulated under each of the 19 different climatologies and their associated atmospheric CO₂ concentrations (hereafter “climatologies”) is compared to each of the observed land cover data sets (Fig. 1). First, the simulated vegetation accuracy differs from one biome to another under the same climatology. For example, the shrub tundra biome simulated using the 2-year climatology agrees well (Kappa statistic = 0.46, Fuzzy Kappa = 0.47 and Nomad index = 0.62) with the GLCC data while other biomes from the same simulation, such as evergreen taiga/montane forest (EvTMF), display relatively poor agreement (Kappa statistic = 0.30, Fuzzy Kappa = 0.30 and Nomad index = 0.38). These differences in simulated accuracy may reflect differences in the model’s ability to simulate particular vegetation types or they may indicate that the input data, in this case the 2-year climatology, fails to capture the long-term mean climate for certain regions.

Second, the simulated vegetation accuracy varies with the length of the climatology used to simulate the vegetation data. For example, under the “BIOME4 vs. GLCC” comparison, when the climatology length is less than 30 years the Kappa statistic, Fuzzy Kappa and the Nomad index values vary substantially from one simulation to the next for most biomes. In contrast, the three statistics tend to vary less from one simulation to the next when the climatology length is greater than 30 years (Fig. 1). This consistency reflects the fact that climatological means spanning longer time periods are less influenced by the addition or subtraction of individual years than are climatological means spanning shorter time periods.

Third, the vegetation simulations run with the 25-, 30- and 35-year climatologies produce simultaneously the highest overall Kappa statistic, Fuzzy Kappa and Nomad index when compared to the three observed data sets (Fig. 2 and Table 6). This relatively good agreement between the simulated and observed vegetation may reflect the general adaptation of regional-scale vegetation to the long-term mean climate of a region. It also suggests that climatologies including 25 or more years are of sufficient duration to capture the interannual and regional climate variations that affect a region’s vegetation (Carter et al., 2001). The overall accuracy of the simulated vegetation supports the continued use of 30-year climatological means, which have been a standard length for climatologies used in vegetation modeling studies.

For the simulations run using 30-year mean climatologies for different time periods, the vegetation simulated using the most recent 30-year climatologies (i.e. 1963–1992 or 1961–1990) produce the highest overall Kappa statistic, Fuzzy Kappa and Nomad index under all three comparisons (Fig. 3). These high values may reflect the influence of recent climate and atmospheric CO₂ concentrations on the GLCC and GLCF observed vegetation data. In
addition, the statistics show that the 1963–1992 30-year climatology can accurately simulate the greatest number of biomes (e.g. 9 out of 16 biomes under the “BIOME4 vs. GLCC” comparison) (Table 7).

The high values for the three statistics in the early part of the 20th century (Fig. 3) may reflect the relaxation of the CRU TS 2.0 monthly climate data for this period towards the 1961–1990 climatology values because of an insufficient number of Asian climate station records (see New et al., 2000). Variations through time in the number of station records included in the CRU TS 2.0 estimated monthly climate data will affect the match between the observed and estimated climate as will limitations of the climate estimation method itself. Errors in the estimated climate will affect the accuracy of the BIOME4-simulated vegetation. In addition, changes in atmospheric CO₂ concentrations through time will influence both observed vegetation and the vegetation simulated by BIOME4 (Diffenbaugh et al., 2003; Harrison and Prentice, 2003). The match between simulated and observed vegetation will also be affected by the ability of BIOME4 to accurately simulate the processes determining the distribution of vegetation.

Although the 5-year climatology produces a high overall Kappa statistic and Fuzzy Kappa under the “BIOME4 vs. PNV” comparison, the values of these statistics are relatively low under the “BIOME4 vs. GLCC” and “BIOME4 vs. GLCF” comparisons (Table 6). Similarly, although the 2-year climatology produces a high Nomad index under the “BIOME4 vs. PNV” and “BIOME4 vs. GLCF” comparisons, it results in the lowest overall Kappa statistic and Fuzzy Kappa under all three comparisons (Fig. 2). Compared to other climatologies, the simulations run using the 2- and 5-year climatologies have the greatest potential to incorrectly simulate the greatest number of

![Fig. 2](image1)

![Fig. 3](image2)
Table 7
Comparison of GLCC biomes with the BIOME4 biomes simulated under the 1963–1992 30-year mean climatology.

<table>
<thead>
<tr>
<th>Reclassified biome types</th>
<th>Kappa statistic</th>
<th>Fuzzy Kappa</th>
<th>Nomad index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Tropical evergreen forest</td>
<td>0.25</td>
<td>0.26</td>
<td>0.30</td>
</tr>
<tr>
<td>2) Tropical semi-deciduous and deciduous forest</td>
<td>0.45</td>
<td>0.48</td>
<td>0.61</td>
</tr>
<tr>
<td>3) Temperate deciduous/conifer forest</td>
<td>0.41</td>
<td>0.43</td>
<td>0.68</td>
</tr>
<tr>
<td>4) Warm mixed forest</td>
<td>0.51</td>
<td>0.54</td>
<td>0.76</td>
</tr>
<tr>
<td>5) Cool/cold mixed forest</td>
<td>0.16</td>
<td>0.17</td>
<td>0.35</td>
</tr>
<tr>
<td>6) Cool conifer forest</td>
<td>0.07</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>7) Evergreen taiga/montane forest</td>
<td>0.38</td>
<td>0.38</td>
<td>0.39</td>
</tr>
<tr>
<td>8) Deciduous taiga/montane forest</td>
<td>0.42</td>
<td>0.40</td>
<td>0.83</td>
</tr>
<tr>
<td>9) Tropical savanna and grassland</td>
<td>0.26</td>
<td>0.28</td>
<td>0.44</td>
</tr>
<tr>
<td>10) Tropical xerophytic shrubland</td>
<td>0.29</td>
<td>0.29</td>
<td>0.65</td>
</tr>
<tr>
<td>11) Temperate xerophytic shrubland</td>
<td>0.40</td>
<td>0.41</td>
<td>0.80</td>
</tr>
<tr>
<td>12) Temperate savanna and grassland</td>
<td>0.36</td>
<td>0.36</td>
<td>0.38</td>
</tr>
<tr>
<td>13) Barren and desert</td>
<td>0.55</td>
<td>0.56</td>
<td>0.89</td>
</tr>
<tr>
<td>14) Steppe tundra</td>
<td>0.50</td>
<td>0.54</td>
<td>0.84</td>
</tr>
<tr>
<td>15) Shrub tundra</td>
<td>0.42</td>
<td>0.42</td>
<td>0.56</td>
</tr>
<tr>
<td>16) Dwarf and prostrate shrub and moss tundra</td>
<td>0.48</td>
<td>0.50</td>
<td>0.58</td>
</tr>
<tr>
<td>Overall statistics</td>
<td>0.40</td>
<td>0.42</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 8
Comparisons among three observed vegetation data sets.

<table>
<thead>
<tr>
<th>Reclassified biome types</th>
<th>GLCC vs. GLCF</th>
<th>GLCC vs. PNV</th>
<th>GLCF vs. PNV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\kappa)</td>
<td>FK</td>
<td>NI</td>
</tr>
<tr>
<td>1) Forest cover</td>
<td>0.69</td>
<td>0.66</td>
<td>0.77</td>
</tr>
<tr>
<td>2) Grassland/tree cover (&lt;10%)</td>
<td>0.56</td>
<td>0.52</td>
<td>0.83</td>
</tr>
<tr>
<td>3) Barren and desert</td>
<td>0.56</td>
<td>0.58</td>
<td>0.47</td>
</tr>
<tr>
<td>Overall statistics</td>
<td>0.61</td>
<td>0.57</td>
<td>0.69</td>
</tr>
</tbody>
</table>

\(\kappa\)—Kappa statistic; FK—Fuzzy Kappa; NI—Nomad index.

individual biomes (Fig. 4), largely because the climate data included in the 2- and 5-year climatologies may not reflect the long-term climate mean of a given area. Assuming that vegetation is responding to conditions that prevail over the long-term as opposed to those of any given year, the accuracy of the vegetation simulated using a 30-year climatology should be better than that simulated using a 2- or 5-year climatology.

3.2. Effect of observed data choice on model accuracy assessment

The choice of what observed vegetation data to use in evaluating model results can significantly affect the assessment of a model’s ability to simulate the overall vegetation pattern for a region. Different observed vegetation data sets may vary significantly in their representation of observed vegetation (Table 8). For example, when the BIOME4 simulations created with the most recent 30-year (1963–1992) climatology are compared with the PNV data, the Kappa statistic \(\kappa = 0.46\) indicates that BIOME4 has a fair ability to simulate potential natural vegetation patterns in continental Asia (Table 9). However, BIOME4’s ability to accurately simulate vegetation is weaker when compared to both the GLCF data \(\kappa = 0.42\) and the GLCC data \(\kappa = 0.40\). As illustrated in Fig. 5, the “BIOME4 vs. PNV” case produces both a higher overall Kappa statistic, a higher overall Fuzzy Kappa and a higher overall Nomad index under all 19 scenarios than either the comparison of “BIOME4 vs. GLCC” or “BIOME4 vs. GLCF” does.

BIOME4 was originally developed to simulate potential natural vegetation and thus it is not surprising that the BIOME4-simulated vegetation better matches the PNV data than it does either the GLCC or GLCF data. The PNV data used in this study were derived from a combination of the DISCover land cover data set (Loveland et al., 2000) and the Haxeltine and Prentice (1996) potential natural vegetation data (Ranumkutty and Foley, 1999). In the PNV data, the vegetation types in regions dominated by human land use were based on the Haxeltine and Prentice (1996) potential natural vegetation data, which in turn were based primarily on the vegetation maps of Melillo et al. (1993). In contrast, both the GLCC and GLCF data were developed completely from remotely sensed AVHRR data.

Table 9
Comparison of PNV biomes with the BIOME4 biomes simulated under the 1963–1992 30-year mean climatology.

<table>
<thead>
<tr>
<th>Reclassified biome types</th>
<th>Kappa statistic</th>
<th>Fuzzy Kappa</th>
<th>Nomad index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Forest cover</td>
<td>0.49</td>
<td>0.45</td>
<td>0.77</td>
</tr>
<tr>
<td>2) Tropical/temperate savanna</td>
<td>0.01</td>
<td>0.01</td>
<td>0.16</td>
</tr>
<tr>
<td>3) Grassland</td>
<td>0.32</td>
<td>0.31</td>
<td>0.41</td>
</tr>
<tr>
<td>4) Tropical/temperate shrubland</td>
<td>0.37</td>
<td>0.37</td>
<td>0.75</td>
</tr>
<tr>
<td>5) Tundra</td>
<td>0.65</td>
<td>0.65</td>
<td>0.78</td>
</tr>
<tr>
<td>6) Barren and desert</td>
<td>0.57</td>
<td>0.58</td>
<td>0.84</td>
</tr>
<tr>
<td>Overall statistics</td>
<td>0.46</td>
<td>0.46</td>
<td>0.73</td>
</tr>
</tbody>
</table>
which were snapshots of the actual vegetation over the period of April 1992 to March 1993. Although we excluded all grid cells with human-altered land cover types in both the “BIOME4 vs. GLCC” and “BIOME4 vs. GLCF” comparisons, other human-induced effects on terrestrial vegetation (e.g. past deforestation, changes in disturbance regimes) may still exist in the reclassified GLCC and GLCF data. In addition, the better match between the BIOME4 and PNV vegetation types (Tables 1 and 2) may contribute to the higher overall statistics in the “BIOME4 vs. PNV” case than in the other two comparisons (Fig. 5).

The extent to which the observed data must be reclassified to be compared with the simulated data will also affect the accuracy assessment of a model to simulate a given biome. For example, the Kappa statistic ($\kappa > 0.55$) indicates that BIOME4 has a good ability to simulate the spatial distribution of barren and desert areas (the only common category classified in all three comparisons) in continental Asia when compared with the GLCC and PNV data (Tables 7 and 9), but a poor ability ($\kappa = 0.35$) when compared with the GLCF data (Table 10). This difference in accuracy is mainly a result of how barren and desert areas are classified in the three observed data sets (Fig. 6a vs. b). These biomes often consist of a small number of continuously distributed grid cells spread over a large area. In contrast, the reclassified barren and desert consists of either “bare desert and sand desert” in the GLCC data or “desert and polar/rock/ice” in the PNV data, both of which were largely based on the “barren or sparsely vegetated” land cover type in IGBP DISCover data set (Ramankutty and Foley, 1999; Loveland et al., 2000) and thus included areas with little vegetation (e.g. the spatial extent of reclassified barren and desert in west China is much larger in both GLCC (Fig. 6b) and PNV (Fig. 6d) than in GLCF data (Fig. 6f)). Because the definitions and attributes of the reclassified barren and desert biome are different in these three data sets, the spatial pattern of the GLCC- and PNV-based barren and desert biomes display better visual agreement with the BIOME4-simulated barren and desert biome than with the GLCF-based biome (Fig. 6).

### 3.3. Effect of map-comparison methods on accuracy assessment

The choice of map-comparison methods will affect the accuracy assessment of a model simulation. Given the same observed and simulated vegetation data, different comparison methods can produce similar judgments of a model’s ability to simulate a given biome. For example, the individual Kappa statistic, Fuzzy Kappa and Nomad index all indicate that BIOME4 can accurately predict temperate deciduous/conifer forest, deciduous taiga/montane forest, temperate savanna and grassland, temperate xerophytic shrubland, steppe tundra and barren/desert (Table 7 and Fig. 6a vs. b), tundra and barren/desert (Table 9 and Fig. 6c vs. d), and forest cover (Table 10 and Fig. 6e vs. f) in continental Asia. These reclassified biomes consist of a large number of continuously distributed grid cells in the two compared maps. In contrast, the three statistics suggest that BIOME4 less accurately simulates cool conifer forest (Table 7 and Fig. 6a vs. b) and tropical/temperate savanna (Table 9 and Fig. 6c vs. d). These biomes often consist of a small number of discretely distributed grid cells spread over a large area.

Different conclusions about a model’s ability to simulate a given biome may also occur when different comparison methods are used for the assessment. For example, for tropical/temperate shrubland in the “BIOME4 vs. PNV” comparison, the low Kappa (0.37) and Fuzzy Kappa (0.37) values suggest an inability of BIOME4 to correctly simulate shrubland in our research area. In contrast, the high Nomad index value (0.75) indicates that BIOME4 has a relatively good ability to simulate shrubland. Fig. 6c and d indicate that the pattern of the BIOME4-simulated shrubland resembles to some degree that of the PNV shrubland. Because the Nomad index assesses pattern agreement based on geographic distance instead of location accuracy (i.e. the correct location of a category), the relatively good agreement of spatial pattern (especially in the mid-latitude areas) between observed and simulated shrubland and a large number of total compared grid cells (see Eq. (4)) result in the high Nomad index. However, because the Kappa and Fuzzy Kappa statistics depend on a pixel-by-pixel comparison that emphasizes

### Table 10

<table>
<thead>
<tr>
<th>Reclassified biome types</th>
<th>Kappa statistic</th>
<th>Fuzzy Kappa</th>
<th>Nomad index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Forest cover</td>
<td>0.52</td>
<td>0.50</td>
<td>0.68</td>
</tr>
<tr>
<td>2) Grassland/percent tree cover</td>
<td>0.33</td>
<td>0.27</td>
<td>0.58</td>
</tr>
<tr>
<td>less than 10%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) Barren and desert</td>
<td>0.35</td>
<td>0.36</td>
<td>0.71</td>
</tr>
<tr>
<td>Overall statistics</td>
<td>0.42</td>
<td>0.41</td>
<td>0.64</td>
</tr>
</tbody>
</table>
Fig. 6. The BIOME4-simulated vegetation for Asia (a, c and e) under the 1963–1992 30-year mean climatology and the observed vegetation from the (b) GLCC, (d) PNV and (f) GLCF data. The white and black areas in (a), (b), (e) and (f) refer to human-modified land and inland water, respectively.

location accuracy, the large disagreement between observed and simulated shrubs in India produces low Kappa and Fuzzy Kappa values.

Although the Kappa and Fuzzy Kappa statistics in this study are approximately equal to each other under all comparisons, the simulated vegetation accuracy assessment is still sensitive to how they are used. For example, Fig. 7 illustrates how the Fuzzy Kappa varies with changes in the neighborhood radius and the fuzzy membership function, two parameters that need to be specified to calculate the Fuzzy Kappa. The Fuzzy Kappa values decrease as the neigh-
atmospheric CO2 concentrations. The BIOME4 simulations run with sitive to the selection of input climatologies and their associated
hood radius of 2 and exponential decay function halving distance of
the “fuzzy membership function,” and category fuzziness values.
must exercise care in choosing the values of neighborhood radius,
the Kappa statistic. Therefore, when using the Fuzzy Kappa to assess
the accuracy of simulated vegetation model, users must exercise care in choosing the values of neighborhood radius, the “fuzzy membership function,” and category fuzziness values. For 0.5-degree data, the Fuzzy Kappa calculated using a neighborhood radius of 2 and exponential decay function halving distance of 2 will greatly underestimate the accuracy of simulated vegetation (Fig. 7).

4. Conclusion

First, the evaluation of global vegetation model accuracy is sen-
tive to the selection of input climatologies and their associated
atmospheric CO2 concentrations. The BIOME4 simulations run with
the most recent 30-year (1963–1992) CRU TS 2.0 climatology and its
associated atmospheric CO2 concentration simultaneously result in
the highest agreement when the simulated biomes are compared
to each of the three land- and tree-cover data sets. In addition,
the simulations driven by 25-, 30- and 35-year climatologies most
accurately simulate the greatest number of individual biomes. Our
results support the use of 30-year (or longer) climatologies (and
their associated atmospheric CO2 concentrations) when using an
equilibrium vegetation model to simulate terrestrial vegetation,
particularly the 30 years that correspond to the time period repre-
sented by the observed vegetation data that will be used to evaluate the
simulated results.

Second, the results reveal that the choice of which observed data
set to compare with the simulated data will affect the assessment
of a vegetation model’s ability to simulate terrestrial vegetation
patterns at global-to-regional scales. The BIOME4-simulated Asian
terrestrial vegetation displays better agreement with the PNV data
than with either the GLCC or GLCF data, because both the BIOME4
simulations and the PNV data share the same goal of describ-
ing potential vegetation determined largely by climate and in the
absence of human activity. However, the agreement between the
BIOME4-biomes and those reclassified from the GLCC and GLCF
data is lower because these two AVHRR-derived data sets are snap-
shots of modern vegetation at a particular time (April 1992 to March
1993) and include the effects of human activities on vegetation
even though we excluded the human-influenced grid cells in our
comparison). Therefore, when choosing observed data for model
evaluation, model users must take into account the nature of both
the model and the observed data. A correct assessment of a model’s
accuracy requires that the attributes and definitions of the sim-
ulated vegetation match that defined or classified from observed data.

Third, the results demonstrate that the use of different map-
comparison methods will affect the evaluation of a model’s accuracy.
Our results reveal that the accuracy assessments of the BIOME4 simulations based on the Kappa statistic, Fuzzy Kappa and Nomad index agree well when the compared vegetation types con-
sist of a large number of continuously distributed grid cells on both
the simulated and observed maps. However, when the patterns of the
simulated and observed biomes are similar but consist of a
large number of grid cells that are not continuously distributed, the
Nomad index tends to indicate better agreement than the Kappa
or Fuzzy Kappa statistics because it stresses pattern agreement
instead of location accuracy. In contrast, the Kappa statistic and
Fuzzy Kappa are better suited for accuracy assessments that empha-
size the location accuracy of the simulated vegetation. In addition,
this study suggests that when using the Fuzzy Kappa to assess a
model’s simulated vegetation accuracy it is essential to consider
the spatial resolution of the compared data. Our results indicate
that a neighborhood radius of 1.0 and exponential decay function
halving distance of 0.5 are good for data with a 0.5 spatial resolu-
tion and that cover a broad area, largely because a larger radius
increases the potential for underestimating a model’s simulation
accuracy by increasing the expected similarity between compared
categories (Fig. 7).

Acknowledgements

Research was supported by NSF grant numbers ATM 9532074
and ATM 9910638, the Department of Geography at the University
of Oregon, and the U.S. Geological Survey Earth Surface Dynamics
Program. We thank Steve Hostetler, Jed Kaplan, Noah S. Diffenbaugh
and one anonymous reviewer for their comments on earlier versions
of this manuscript. Contributions: The overall research design
was conceived by G. Tang, S. Shafer, and P. Bartlein, with input from J.
Holman. J. Holman developed the Nomad index and contributed to
the analysis of map-comparison methods. G. Tang prepared the ini-
tial draft of the manuscript, tables, and figures, with contributions
from all authors.

References

Allen, M., Raper, S., Mitchell, J., 2001. Climate change—uncertainty in the IPCC’s third
Barrett, D.J., Galbally, I.E., Graetz, R.D., 2001. Quantifying uncertainty in estimates of
C emissions from aboveground biomass due to historic land-use change to
Andrew, A.A., Bartlein, P.J., Christensen, T.R., Cramer, W., Kaplan, J.O., Lohzkin, A.J.,
Matsuoka, Y., Murray, D.F., McGuire, A.D., Razzhivin, V.Y., Ritchie, J.C., Smith, B., Walker, D.A.,
Gajewski, K., Wolf, V., Holmgren, B.H., Igarashi, Y., Kremenetski, K., Paus, A., Pisacicic,
MFJ, Volka, V.S., 2003. Climate change and Arctic ecosystems I: Vegetation changes north of 55
between the last glacial maximum, mid-Holocene and present. J. Geophys. Res. 108 (D19),
Bolliger, J., Kienast, F., Bugmann, H., 2000. Comparing models for tree distributions:
20, 37–46.
Congalton, R.G., Green, K., 1999. Assessing the Accuracy of Remotely Sensed Data:
Model. 47, 199–215.
global 1 km data set of percent tree cover derived from remote sensing. Glob.
Change Biol. 6, 247–254.


