

**ENHANCING THE EFFICIENCY OF TERRESTRIAL BIOSPHERE MODEL  
SIMULATIONS BY REDUCING THE REDUNDANCY  
IN GLOBAL FORCING DATA SETS**

**Author:** Axel Kleidon  
Department of Geography and  
Earth System Science Interdisciplinary Center  
2181 Lefrak Hall  
University of Maryland  
College Park, MD 20742  
USA

e-mail: akleidon@umd.edu  
phone: 301-405-3203  
fax: 301-314-9299

Working Paper  
“02-Cluster.pdf”

deposited on the Digital Repository at the University of Maryland  
<http://drum.umd.edu>

1/10/05

## ABSTRACT

Data sets of climatic variables and other geographic characteristics are becoming available in increasingly higher resolutions, resulting in substantial computing burdens for simulation models of the terrestrial biosphere. But by how much do higher resolutions of forcing data actually contribute to higher accuracy in model predictions? I investigated this question using the Cramer-Leemans climatology as an example for a high resolution forcing data set and a model of net primary productivity (*NPP*). I first used cluster analysis to reduce the complete grid of the climatology to a few grid points, each representative of regions with similar values. A global map of *NPP* was reconstructed by using the simulated values of the representative grid points for the respective regions. I then compared the reconstructed map of *NPP* to the one obtained from all grid points. The results show that a high accuracy in simulating the high resolution pattern and magnitude can be achieved by only considering a comparatively small subset of representative grid points. What this suggests is that, while high resolution data sets provide the necessary means to determine the typical regions, they do not add much accuracy to the overall outcome of model simulations because they contain many grid points with similar values. By reducing this redundancy, the methodology used here allows model simulations to be considerably more computing-time efficient while still retaining the accuracy in predicted quantities.

## INTRODUCTION

Numerical simulation models are often used to predict properties of terrestrial ecosystems, for the present-day and for scenarios of global change (e.g. VEMAP Members 1995, Heimann et al. 1998, Cramer et al. 1999). In order to achieve a high level of predictive skills, the model's spatial resolution is often high (1 degree latitude or finer), requiring high-resolution data sets of forcing variables such as temperature and precipitation. However, large regions across the world often show similar environments, resulting in similar composition and functioning of ecosystems. The similarity in composition is expressed by the concept of vegetation types or biomes while the functional similarity manifests itself in the convergence in plant functioning (Reich et al. 1997). Therefore we may ask how much information high-resolution data sets contain, that is, how many

distinct groups of similar environments are described by these data sets. Extracting such distinct groups will provide a means to considerably increase the computing time efficiency of simulation models, such as Dynamic Global Vegetation Models (DGVMs, e.g. Foley et al 1996, Neilson and Running 1996), without losing much of its predictive accuracy.

I will describe such a methodology and its power in this study. The idea is to first extract similar regions from the data set by using cluster analysis (e.g. Späth, 1985), then to run the simulation model only for one representative point for each region, and finally to reassemble a high-resolution map of predicted quantities by using the simulated value for the representative points. A schematic diagram of this setup is shown in Figure 1. As an example for this method, I used the high-resolution, global climatology data set of Cramer and Leemans (Cramer and Leemans, pers. comm., updated version of Leemans and Cramer, 1991) and extracted regions with similar climatic characteristics. For different numbers of typical regions, I applied a simulation model of net primary productivity (*NPP*) to the representative points, and reassembled the global map of *NPP*. By comparing the reassembled map to one from a simulation for all grid points, I then investigated how the chosen number of regions for this method affect the accuracy of the reassembled map of *NPP*.

In the next section I explain the details of the methodology as well as the climatology, the simulation model for *NPP*, the overall setup and the evaluation strategy. The results are presented in form of global maps of *NPP*, which were reassembled from different number of regions, and a measure of the accuracy of each of the maps. I close with a discussion on the possible applications and limitations of this method.

## **METHODS**

### *Description of the Methodology*

The methodology is based on cluster analysis (e.g. Späth, 1985). A cluster is defined as an object of data points with similar values. The similarity of data points is measured by their Euclidian distances to the cluster center. I use the “minimal distance method” to obtain the clusters as described by Späth (1985). Given a fixed number of clusters, this method iteratively

minimizes the variance within a cluster. The quality of clustering can be described by the ratio of the between-cluster-variance to the within-cluster-variance. A ratio greater than one implies that the cluster points are separated; the greater the ratio, the better the separation. This ratio will be referred to as “separation” in the following sections.

The set up was as follows: First, cluster analysis was performed for 2 clusters. The number of clusters was subsequently increased until the separation between the clusters exceeded a value of 2. Then, each of the clusters obtained at this threshold was treated individually and clustered again, using the same procedure. For each of the clusters, a representative grid point was determined by taking the grid point whose climate is closest to the mean climate of all cluster members. The simulation model was run for the representative grid points only. The simulated value for the representative grid points was then used for all grid points of the clusters in order to reassemble global maps of simulated values.

### *Description of the Climatology*

I used the Cramer-Leemans (C&L) climatology (Cramer and Leemans, pers. comm., updated version of Leemans and Cramer, 1991), which is frequently used in global modelling studies (e.g. Heimann et al. 1998, Cramer et al. 1999). The C&L climatology consists of five data sets of monthly mean values for precipitation, number of rainy days, daily temperature, daily temperature range and cloudiness. It uses a terrestrial grid of 0.5 degree resolution, resulting in a total of 62,483 grid points. The annual mean characteristics of this climatology are shown in Figure 2 in terms of temperature and precipitation.

I first converted cloudiness into solar radiation according to Linacre (1968) and Prentice et al. (1993) in order to convert it into a more meaningful variable representative of surface conditions. Then each of the 5 variables was scaled between 0 and 100, with the limits representing the minimum and maximum of each variable of the whole data set respectively. This scaling was done in order to make different climatic variables comparable, and by scaling them to the same range an equal weight is put on each variable in the cluster analysis. Note that this rescaling is only done for computing the similar regions from the climatology. The methodology described above was applied to this climatology by treating each of the 62,483 grid points as one data point of 60

dimensions (5 variables with 12 monthly values each). In this way, the seasonal characteristics are explicitly included in the cluster analysis. The iterative clustering process, as described above, was done up to five times in total. In the first iteration, 6 clusters, or regions, were obtained. These 6 clusters were partitioned into 55 clusters in the second iteration, 506 in the third, 2976 in the fourth and 13371 clusters in the fifth, last iteration.

### *Description of the Simulation Model*

In order to investigate the effect of different numbers of clusters on the simulated results, I used a model for net primary productivity (*NPP*) as described in Kleidon and Heimann (1998). This model uses a light use efficiency formulation for computing *NPP* based on Monsi and Saeki (1957) and Monteith (1977) in combination with a bucket-type soil hydrology model to compute water stress (Prentice et al. 1993). Water stress is computed from the ratio of actual to potential evapotranspiration. Actual evapotranspiration is calculated as the minimum of supply, which depends on soil wetness, and the atmospheric demand (Federer 1982). The atmospheric demand is approximated by the equilibrium evapotranspiration rate of McNaughton and Jarvis (1983). The model uses daily time step. For each day, the mean monthly forcing was used. The number of rainy days and the daily temperature range was not used by the model.

### *Evaluation Strategy*

In order to quantify the power of this methodology, the reconstructed maps of *NPP* were compared to the one obtained from simulating all grid points of the climatology. I will use the term “exact” in reference to the results of the simulation with all grid points considered. The simulation model was run for the representative grid points of the clusters determined at different stages of the clustering process, that is, for 6, 55, 506, 2976, and 13371 clusters, representing 0.01%, 0.09%, 0.81%, 4.76%, and 21.40% of the total number of grid points respectively. Each of the reconstructed maps of annual *NPP* are shown as well as the exact map in the results section.

The accuracy of the method was assessed in terms of relative differences in the simulated annual *NPP* between the reconstructed map ( $NPP_R$ ) and the exact map ( $NPP_E$ ). On the grid point level, a mean “local error” was quantified by summing up the absolute differences between  $NPP_R$

and  $NPP_E$  at all grid points, weighted by their respective area  $A$ :

$$\Delta_{LOCAL} = \frac{\sum_{\text{all gridpoints}} |NPP_R - NPP_E| \cdot A}{\sum_{\text{all gridpoints}} NPP_E \cdot A} \quad (1)$$

On a larger scale, I quantified a mean “latitudinal error” by first computing latitudinal sums of  $NPP$  and then summing up the differences:

$$\Delta_{LAT} = \frac{\sum_{\text{all latitudinal bands}} \left| \sum_{\text{latitudinal band}} NPP_R - \sum_{\text{latitudinal band}} NPP_E \right| \cdot A}{\sum_{\text{all gridpoints}} NPP_E \cdot A} \quad (2)$$

Finally, I quantified a “global error” by taking the differences of global sums of annual  $NPP$  between both simulations:

$$\Delta_{GLOBAL} = \frac{\sum_{\text{all gridpoints}} NPP_R \cdot A - \sum_{\text{all gridpoints}} NPP_E \cdot A}{\sum_{\text{all gridpoints}} NPP_E \cdot A} \quad (3)$$

These measures for the error of the described method were calculated for each of the five clustering setups.

## RESULTS AND DISCUSSION

The reconstructed maps of annual  $NPP$  are shown in Figure 3a-e as well as the map obtained by simulating all grid points (exact map, Figure 3f). The first map, Figure 3a, uses only 6 regions/representative grid points and obviously shows a coarse pattern. The boundaries of the regions roughly correspond to what one would refer to arctic, temperate, arid, semiarid, and tropical humid regions. While the local error is high (Figure 4), the global  $NPP$  of this reconstructed map is already within 5% of the exact value. The boundaries between the regions clearly diminish with increasing number of regions, and no substantial improvement can be seen in Figures 3d and 3e compared to 3f. The simulated value of global  $NPP$  also does not improve above more than 506

---

regions (Figure 4), while some mismatch still exists at the grid point level.

Capturing the large-scale simulated pattern by simulating only a comparatively small number of representative grid points – Figure 3c represents less than 1% of all grid points – is impressive. This result, however, is not surprising. As mentioned in the introduction, the similarity of regions in respect to their climatic environment gave rise to climate classifications. Not only is the large-scale pattern well reproduced by a comparatively few representative grid points, but also the global mean. The local values, however, deviate more. This discrepancy in error reduction on local versus global scale is an inherent characteristic of the methodology. The representative grid points typically lie close to the center of their regions. This means, that the grid points surrounding the representatives typically show higher values in one direction towards the boundary of the region and lower values towards the opposite direction. When integrated over the whole region, these deviations tend to cancel out each other, thus leading to better predictive skills on the larger scale. This effect is also reflected in the latitudinal error (Figure 4), which is always in between the local and global error.

There are surely cases where individual grid points need to explicitly be simulated rather than a representative grid point. For instance, models in which individual grid points interact with each other, such as coupled land surface-climate models, explicit modelling of all grid points is required. However, in this case, global data sets such as a climatology will mainly serve as a means of model testing rather than as model input. Using representative grid points may also not work in transient model simulations, in which the forcing variables change in time, such as scenarios of global warming. In such a case, nevertheless, the representative grid points may still be characteristic of transient response, and a predicted shift in vegetation zones, for instance, may be captured by the shift in cluster boundaries. This aspect would need further investigations. The methodology can also be extended to the case where more forcing data sets are needed for a model, for instance, data sets on soil texture or vegetation type. These data sets should then be included into the clustering process. Since both of the examples stated correlate with climatic features, the methodology described here should still be applicable. The clustering procedure can be improved as well, for instance, by first using a principal component analysis in order to remove correlations

---

among the forcing variables (e.g. Fowell and Fowell, 1993). This would improve the method in that either less representative grid points are needed or that the accuracy of the method increases.

## CONCLUSION

Global, high resolution data sets, such as a climatology, contain large regions with similar characteristics, thus containing a certain degree of redundancy. By reducing this redundancy with the methodology established here, I showed that simulations with a productivity model can be executed considerably more time efficient without much loss in accuracy. For example, by only considering merely 1% of the grid of a global climatology, a reasonably adequate map of *NPP* can be produced with a global sum of *NPP* indistinguishable from a full model simulation. The high resolution data sets are nevertheless an integral part of the methodology itself in that they provide the necessary means to extract and define regions with similar characteristics. This overall setup is a promising way for making model simulations of the terrestrial biosphere, sensitivity studies and model tuning much more time efficient.

## ACKNOWLEDGEMENTS

The author would like to acknowledge the financial support of the Alexander-von-Humboldt Foundation through a Feodor Lynen Fellowship. Partial support was provided by NASA through a grant from the EOS program (grant number NAS5-31726). The classifications of the Cramer-Leemans climatology used in this paper are available from the author upon request.

## REFERENCES

Cramer W, Kicklighter D W, Bondeau A, Moore III B, Churkina G, Nemry B, Ruimy A, Schloss A L, Kaduk J, et al. (1999). Comparing global models of terrestrial net primary productivity (NPP): overview and key results. *Global Change Biology*, **5** (Suppl. 1): 1-15.

Federer CA (1982) Transpirational supply and demand: plant, soil, and atmospheric effects evaluated by simulation. *Water Resources Research* **18**: 355-362.



---

Foley JA, Prentice IC, Ramankutty N, Levis S, Pollard D, Sitch S, Haxeltine A (1996) An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. *Global Biogeochemical Cycles*, **10**(4): 603-628.

Fowell, RG, Fowell M-YC (1993) Climate zones of the conterminous United States defined using cluster analysis. *Journal of Climate* **6**(11): 2103-2135.

Heimann M, Esser G, Haxeltine A, Kaduk J, Kicklighter D W, Knorr W, Kohlmaier G H, McGuire A D, Melillo J, Moore III B, Otto R D, Prentice I C, Sauf W, Schloss A, Sitch S, Wittenberg U, Würth G (1998) Evaluation of terrestrial carbon cycle models through simulations of the seasonal cycle of atmospheric CO<sub>2</sub>: First results of a model intercomparison study. *Global Biogeochemical Cycles* **12**: 1-24.

Kleidon A, Heimann M (1998) A Method of Determining Rooting Depth from a Terrestrial Biosphere Model and its Impacts on the Global Water- and Carbon Cycle. *Global Change Biology*, **4**(3): 275-286.

Leemans R, Cramer W (1991) *The IIASA Climate Database for Mean Monthly Values of Temperature, Precipitation and Cloudiness on a Terrestrial Grid*. RR-91-18, Institute for Applied Systems Analysis, Laxenburg/Austria.

McNaughton KG, Jarvis PG (1983) Predicting the effects of vegetation change on transpiration and evaporation. *Water Deficit and Plant Growth*. T. T. Kozlowski. New York, Academic Press: 1-47.

Monsi M, Saeki T (1953) Über den Lichtfaktor in den Pflanzengesellschaften und seine Bedeutung für die Stoffproduktion. *Jap. J. Bot.*, **14**: 22-52.

Monteith JL (1977) Climate and the efficiency of crop production in Britain. *Phil. Trans. R. Soc. Lond. B*, **281**: 277-294.

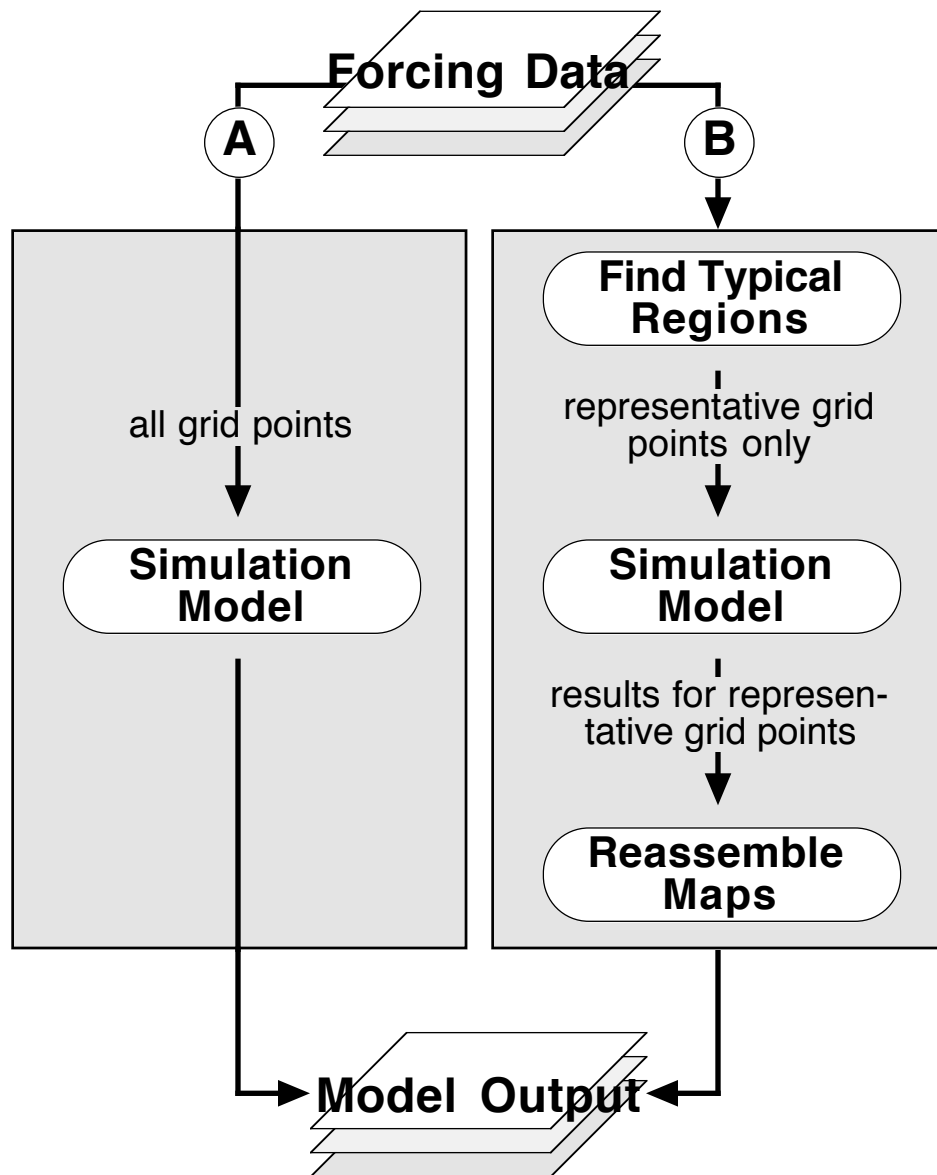
Neilson R, Running SW (1996). Global dynamic vegetation modelling: coupling biogeochemistry and biogeography models. *in: Walker B, Steffen W, eds. Global Change and Terrestrial Ecosystems*. Cambridge University Press, Cambridge, 451 -465.

Prentice, IC, Sykes MT, Cramer W (1993) A simulation model for the transient effects of climate change on forest landscapes. *Ecol. Modelling*, **65**, 51-70.

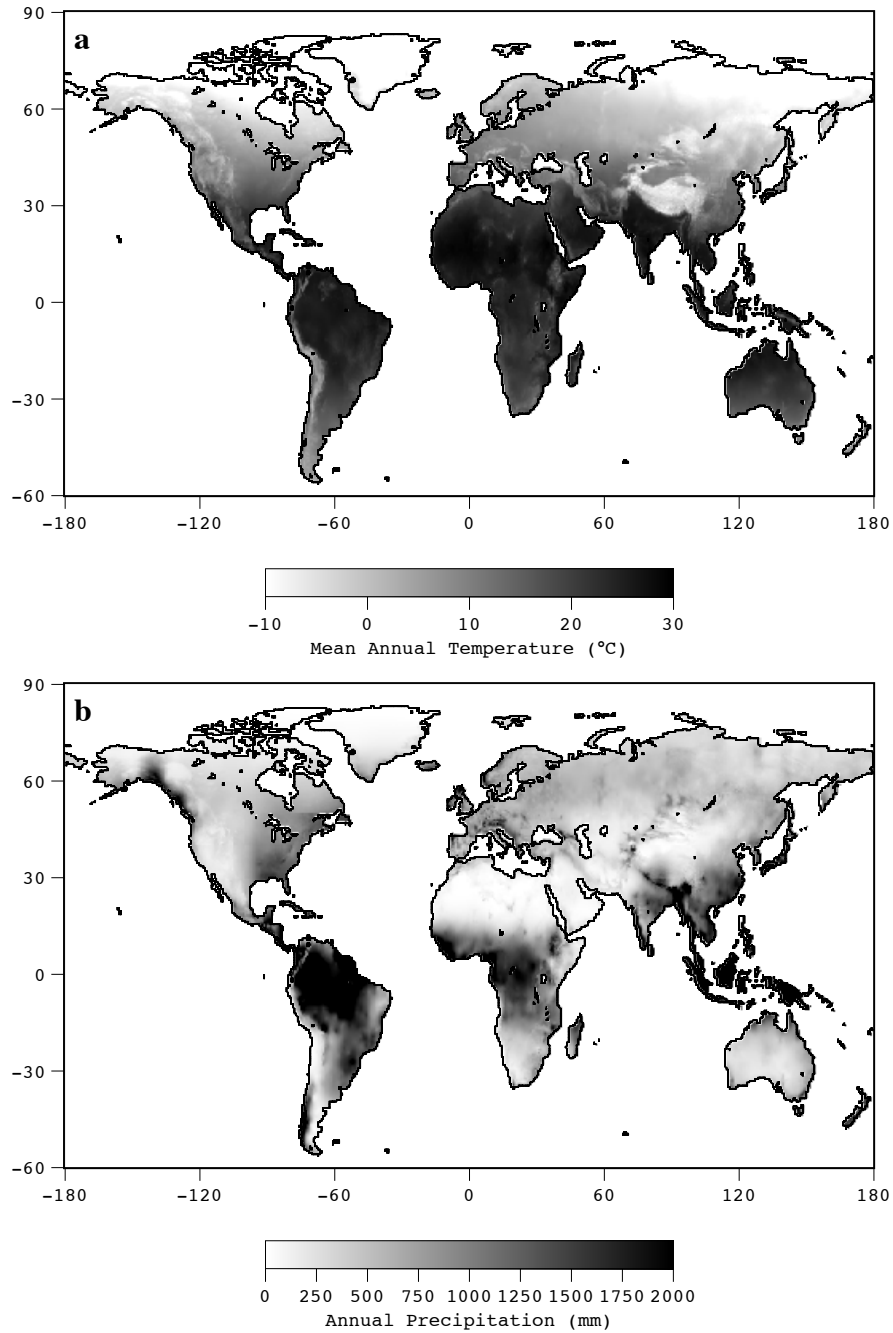
Reich P B, Walters M B, Ellsworth D S (1997) From tropics to tundra: Global convergence in plant functioning. *Proceedings of the National Academy of Sciences USA* **94**: 13730-13734.

Späth H (1985) *Cluster dissection and analysis: theory, FORTRAN programs, examples*. Ellis Horwood Limited, Chichester, England.

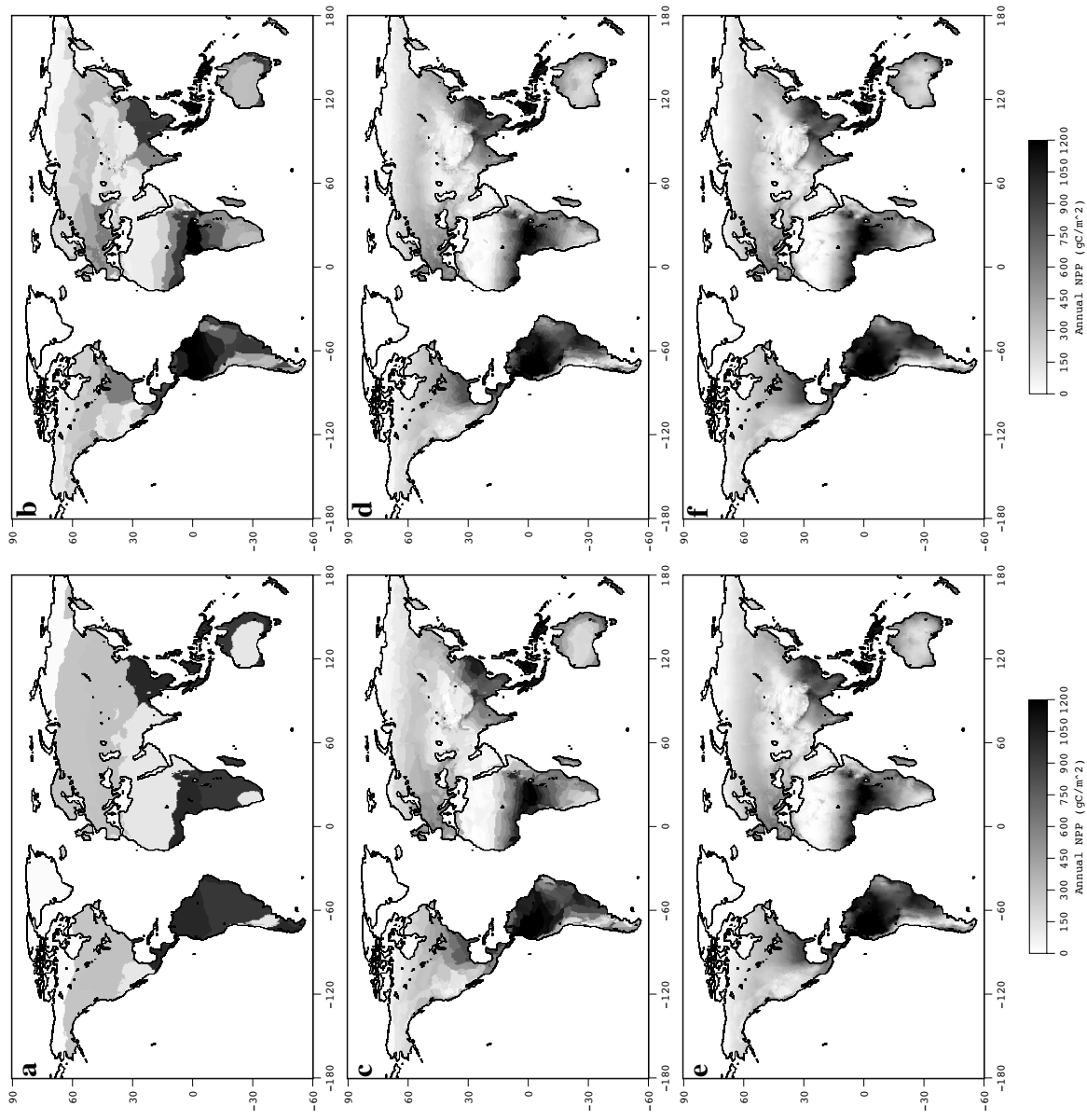
VEMAP members (1995) Vegetation/ecosystem modeling and analysis project: comparing biogeography and biogeochemistry models in a continental-scale study of terrestrial ecosystem responses to climate change and CO<sub>2</sub> doubling. *Global Biogeochemical Cycles* **4**: 407-437.



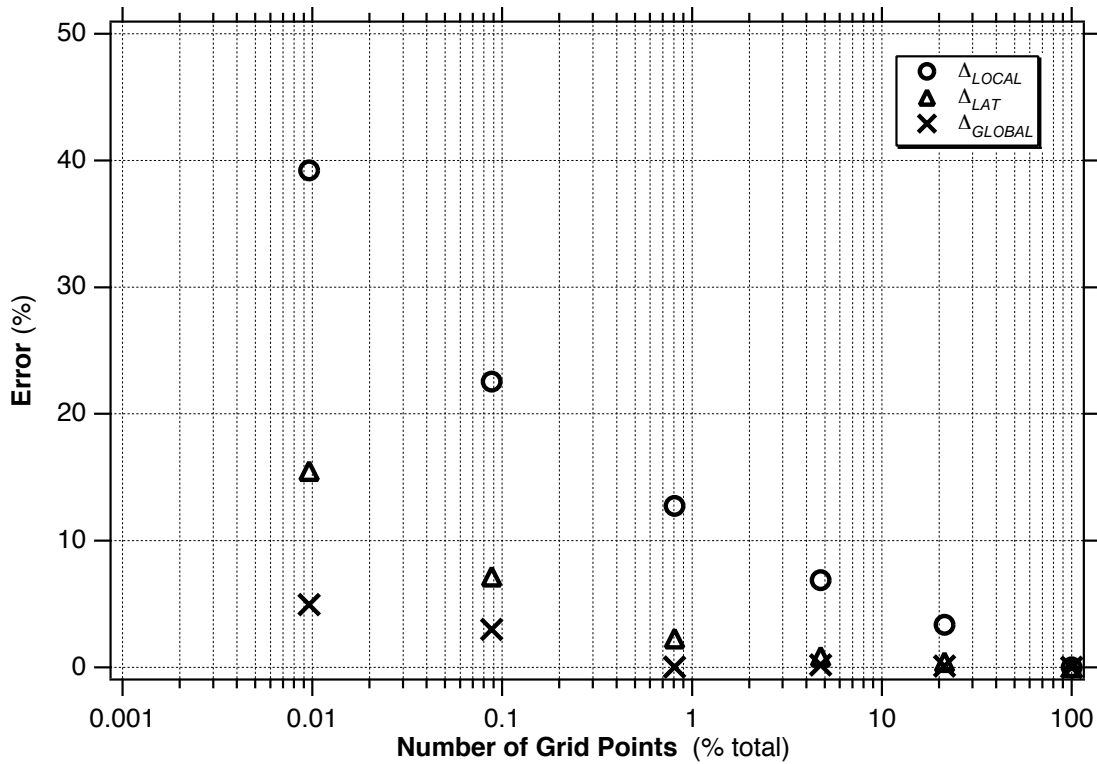
**Figure 1:** Schematic diagram of the conventional model setup (pathway A) and the novel approach presented here (B).



**Figure 2:** The Cramer-Leemans climatology shown in terms of (a) annual mean temperature and (b) annual total precipitation.



**Figure 3:** Maps of annual net primary productivity (*NPP*) reconstructed from the values of (a) 6, (b) 55, (c) 506, (d) 2976 and (e) 13371 representative grid points and (f) obtained from using all 62483 grid points of the climatology.



**Figure 4:** Local, latitudinal, and global error in simulated annual *NPP* according to equations (1) - (3) versus the number of representative grid points used to reconstruct global maps. Both errors are expressed as relative values in relation to the global *NPP* of the simulation with all grid points considered. The number of representative grid points is expressed in relation to the total number of grid points in the climatology.