Using PCA and Global Smoothing to Explore Differences between Global Vegetation Models

Lindström, Johan; Ahlström, Anders; Blom, Emma

Published in:
Proceedings of the 58th World Statistics Congress of the International Statistical Institute (ISI 2011)

2011

Using PCA and Global Smoothing to Explore Differences between Global Vegetation Models

Lindström, Johan
Lund University, Centre for Mathematical Sciences
Matematikcentrum, Box 118, 221 00 Lund, Sweden
E-mail: Johan.Lindstrom@matstat.lu.se

Ahlström, Anders
Lund University, Department of Earth and Ecosystem Sciences
Sölvegatan 12, 223 62 Lund, Sweden
E-mail: Anders.Ahlstrom@nateko.lu.se

Blom, Emma
Lund University, Centre for Mathematical Sciences
Matematikcentrum, Box 118, 22100 Lund, Sweden
E-mail: —

Abstract

A common method for comparing the result of different global circulation models (GCMs) under different emission scenarios is to study global climate response variables, such as mean temperature. An interesting alternative measure of climate sensitivity is to study the biosphere's response to the different climate scenarios. The Lund-Postdam-Jena (LPJ) global vegetation model and its extension LPJ-GUESS is a dynamic global vegetation model that can be coupled to GCMs and used to explore the effect of varying climates on vegetation and carbon uptake.

Using the output from different GCMs under different emission scenarios LPJ-GUESS can be used to generate global vegetation and carbon uptake patterns that are specific to each forcing climate scenario. We investigate if important regional and global differences exist between the vegetation patterns from different GCMs and emission scenarios. An important question is if potential differences are primarily due to the different emission scenarios or to the different GCMs.

In order for us to carry out the above analysis we need to both reduce the noise in the LPJ-GUESS predictions and reduce the vast amount of data. To accomplish both these goals we compute smooth principal components. A problem when computing the PCA and the smoothing is that LPJ-GUESS output is generated on a regular longitude-latitude grid, implying that both the size and distance between grid cells vary. To handle this irregular data on a sphere we use a Gaussian Markov random field (GMRF) approximation of Thin Plate Splines (TPS) that generalises the TPS to general manifolds (such as a sphere). The well known computational advantages of GMRFs greatly aids the analysis, given the large amount of data obtained from LPJ-GUESS.

Introduction

An important question in climate research is the potential effects of climate change. A popular measure when comparing the effects of different global circulation models (GCMs) under different emission scenarios is the increase in the global mean temperature (Boer & Yu, 2003). Although a simple summary measure, this and other global summary statistics might not capture important regional variations (see Boer & Yu, 2003). Further, studying only changes in climate variables — either at the global or local scale — does not provide information regarding the biosphere's response to those changes.

To evaluate the biosphere's potential respons we have used the output from GCMs as driver for a global vegetation model (LPJ-GUESS, see Smith et al., 2001, for details). The output from LPJ-GUESS under different forcing then gives an indication of how the biosphere would respond to different climate scenarios.

The overarching question is if climate change leads to an increase or decrease in the biosphere's uptake of CO_2. A decrease in the biosphere's CO_2 uptake would adversely effect atmospheric CO_2-levels; causing a potentially serious feedback effects for global warming (see Cox et al., 2000).

The goal of this initial analysis is to investigate the differences between the biosphere's response for different combinations of GCMs and greenhouse gas (GHG) emission levels. Specifically we are interested in how much of the spatial variability that is due to differences in emissions and how much is due to differences in the GCMs.

The data

For this study we used the output from four different GCMs (CM4, ECHAM5, CCSM3, and HADCM3; see Marti et al., 2006; Roeckner et al., 2003; Collins et al., 2006; Gordon et al., 2000, respectively). Each GCM was run under 3 of the Intergovernmental Panel on Climate Change's (IPCC's) emission scenarios (A1B, B1, and A2; see Nakicenovic & Swart, 2000) for a total of 12 possible future climates. Here A2 has ever increasing GHG emission, while both A1B and B1 have emissions that initially increase before peaking and declining, with B1 having the lowest emissions. All simulations were initially spun-up and forced over the historical period (1901–2000) with the CRU ts 3.0 dataset (Mitchell & Jones, 2005). At 2001 the GCM-scenario data was superimposed on the CRU 1961–1990 climatology using the delta-change approach.

Given the 12 climate scenarios we used LPJ-GUESS to simulate carbon fluxes (see below) due to terrestrial vegetation. The simulations where carried out for a regular longitude/latitude grid with a 0.5° resolution, giving a total of $\sim 60'000$ grid cells containing vegetation. The flux in each grid cell was aggregated to a future 30 year average over the years 2071–2100; giving us the average carbon flux over 30 years per m^2 for each cell.

The resulting values are very noise with big differences between neighbouring cells. The noise is essentially due to Monte-Carlo type errors, and most of the variability could have been reduced by a longer run of LPG-GUESS. However, the net ecosystem exchange (NEE; see (1) below) is a fine balance between large fluxes of uptake and release of carbon in the ecosystem, and to achieve a noteworthy reduction in noise would require a very considerable increases in computational time.

Carbon flux

The carbon fluxes simulated by LPJ-GUESS represents the amount of carbon either released by vegetation (positive values) or sequestrated (negative values); see e.g. Fig. 2.

The major components of the ecosystems carbon cycle consist of: gross primary production (GPP), i.e. carbon that is sequestrated, mainly through photosynthesis; autotrophic respiration (Ra), carbon released due to the plants' metabolism; heterotrophic respiration (Rh), the release of carbon by microbes and other organisms that consume dead and decaying biomass; and wildfires (F), which constitutes an additional process that releases a sizable fraction of carbon each year. The resulting NEE is defined as

$$\text{NEE} = \text{Ra} + \text{Rh} + \text{F} - \text{GPP}, \tag{1}$$

and provides a measure of the net carbon exchanged between the ecosystem and the atmosphere.

Theory and Analysis

The main idea in the analysis is to use smoothing splines followed by a principal component analysis (PCA) to suppress the noise and extract common spatial patterns for the NEE fields. By regressing the original NEE fields on the PCAs and applying an ANOVA type analysis to the regression coefficients we can then separate the variability into common, GCM, emission scenario, and individual effects.

First we need some notation; let $y_{ij}(s_k)$ denote the NEE from the i^{th} GCM and j^{th} emissions scenario in the grid square centred at s_k and Y_{ij} is a column vector containing $\{y_{ij}(s_k)\}_{k=1}^{N}$. There is $n_i = 4$ different GCMs and $n_j = 3$ emission scenarios. Finally the area of each grid square is denoted a_k.

Spline smoothing using Gaussian Markov random fields

Wahba (1981) noted that a natural extension of Duchon splines (1976) to the sphere can be formulated as solutions to the stochastic partial differential equation (SPDE) $\Delta^{m/2}x(s) = \mathcal{W}$, where Δ is the Laplacian, $\mathcal{W}$ is Gaussian white noise and $m = 2$ essentially gives thin place splines. Several authors (e.g. Kimeldorf & Wahba, 1970; Nychka, 2000) have already noted the similarities between spline smoothing and (Gaussian) spatial-processes. Further Whittle (1954, 1963) pointed out that Gaussian fields with Matérn covariance are solutions to the SPDE $(\varkappa^2 - \Delta)^{m/2} = \mathcal{W}$, where the range is $\propto 1/\varkappa$. This equality has been used by Lindgren et al. (2011) to construct Gaussian Markov random fields (GMRFs) that approximate fields with Matérn covariances. The above can be used to create GMRFs that approximate the splines proposed by Wahba (1981).

Starting with the grid centres we triangulate the sphere, adding a few points over the oceans and close to the poles. Following Lindgren et al. (2011) we then create a latent GMRF on the triangulation. Seeing the NEE fields as noisy observations of a smooth latent field the model becomes

$$(2) \quad X \in N\left(0, \left(\tau^2 Q\right)^{-1}\right) \qquad\qquad Y|X \in N\left(AX, \sigma^2 \mathbf{\Sigma}\right).$$

Here Q is the precision of an intrinsic GMRF (Rue & Held, 2005, chap. 3) obtained by taking $\varkappa = 0$ and $m = 2$ in the SPDE, A is an observation matrix picking out the points in the triangulation that correspondes to grid cells, and $\mathbf{\Sigma}$ is a diagonal covariance matrix that accounts for the varying area of the grid cells. The the diagonal elements of Sigma are $\mathbf{\Sigma}_{kk} = a_k^{-1} / \left(N^{-1} \sum_i a_i^{-1}\right)$.

The spline smoothing of Y is now obtained through the conditional expectation

$$(3) \quad E\left(X|Y; \tau^2, \sigma^2\right) = \left(\tau^2 Q + A^\top (\sigma^2 \mathbf{\Sigma})^{-1} A\right)^{-1} \left(A^\top (\sigma^2 \mathbf{\Sigma})^{-1} Y\right) = \left(\lambda Q + A^\top \mathbf{\Sigma}^{-1} A\right)^{-1} \left(A^\top \mathbf{\Sigma}^{-1} Y\right),$$

where $\lambda = \tau^2 \sigma^2$, with larger values of λ giving smoother reconstructions. Thus the smooth fields, as a function of λ, are $\boldsymbol{\mathcal{Y}} = A\ E\left(X|Y; \lambda\right)$; see e.g. Fig. 2.

Weighted principal component analysis

Having obtained spatially smooth(er) NEE fields a weighted PCA (wPCA) is utilised to determine the major patterns of spatial variability. The weighting is due to the area of the grid cells (Quadrelli & Wallace, 2004). The wPCA is computed for standardised data, $\widehat{\boldsymbol{\mathcal{Y}}}_{ij} = (\boldsymbol{\mathcal{Y}}_{ij} - 1\mu_{ij})/\sigma_{ij}$, where 1 is a vector of ones, $\mu_{ij} = (1^\top \mathbf{\Sigma}^{-1} 1)^{-1} (1^\top \mathbf{\Sigma}^{-1} \boldsymbol{\mathcal{Y}}_{ij})$, and $\sigma_{ij} = \left((\boldsymbol{\mathcal{Y}}_{ij} - 1\mu_{ij})^\top \mathbf{\Sigma}^{-1} (\boldsymbol{\mathcal{Y}}_{ij} - 1\mu_{ij})/N\right)^{1/2}$. Given standardised data the wPCA is computed using a singular value decomposition (Aguiar & Moura, 2003)

$$(4) \quad U = \mathbf{\Sigma}^{1/2} \widetilde{U}, \qquad \text{with } \mathbf{\Sigma}^{-1/2} \widehat{\boldsymbol{\mathcal{Y}}}_{ij} = \widetilde{U} S \widetilde{V}^\top, \qquad \text{where } \widetilde{U}^\top \widetilde{U} = I \text{ and } \widetilde{V}^\top \widetilde{V} = I.$$

Additionally U is orthonormal under the weighted scalar product $U^\top \mathbf{\Sigma}^{-1} U = I$.

Cross-validation to determine smoothing

To determine the optimal value of λ a leave one out cross-validation is used. For each of the 12 fields the remaining 11 are used to compute a wPCA of the smooths, $\boldsymbol{\mathcal{Y}}$. The weighted least squares (WLS) of the left out original field against a regression matrix consisting of an intercept plus the leading principal components (PCs) — i.e. the first columns of U — is then computed. The total weighted RMSE of the WLS residuals is then minimised w.r.t. λ. Using 5 or more components gives $\lambda \approx 7 \cdot 10^{-6}$.

Weighted regression and ANOVA separation

The leading M PCs — given a smoothing based on the optimal λ — are now used to create a regression basis $\widehat{\mathrm{U}} = \begin{bmatrix} 1 & \mathrm{U}_{\cdot,1\ldots\mathrm{M}} \end{bmatrix}$, and regression coefficients for each field are obtained through WLS as

$$\gamma_{ij} = (\widehat{\mathrm{U}}^\top \boldsymbol{\Sigma}^{-1} \widehat{\mathrm{U}})^{-1} (\widehat{\mathrm{U}}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mathcal{Y}}_{ij}) \qquad \text{with residuals} \qquad \mathrm{Y}_{ij} - \widehat{\mathrm{U}}\gamma_{ij} = \boldsymbol{\varepsilon}_{ij}. \tag{5}$$

The regression coefficients are then decomposed into common, GCM, emission, and individual (or interaction) terms using ANOVA

$$\gamma_{ij} = c + \alpha_i + \beta_j + \phi_{ij}. \tag{6}$$

Here α_i, β_j, and γ_{ij} denote deviations from the common effect, and they all sum to zero. Using (6) each field can be divided into a common contribution ($\widehat{\mathrm{U}}c$), contributions from the GCMs ($\widehat{\mathrm{U}}\alpha_i$) and scenarios ($\widehat{\mathrm{U}}\beta_j$), individual parts ($\widehat{\mathrm{U}}\phi_{ij}$), and residuals ($\boldsymbol{\varepsilon}_{ij}$). Further, the (weighted) sum of squares for the total and for each component can be computed as

$$\begin{aligned} &\mathrm{SS}_{\mathrm{tot}} = \sum_{ij} (\mathrm{Y}_{ij} - \mu)^\top \Sigma^{-1} (\mathrm{Y}_{ij} - \mu), \quad \mathrm{SS}_{\mathrm{com}} = n_i n_j \left(\widehat{\mathrm{U}}c - \mu\right)^\top \Sigma^{-1} \left(\widehat{\mathrm{U}}c - \mu\right), \\ &\mathrm{SS}_{\mathrm{gcm}} = n_j \sum_i \alpha_i^\top \widehat{\mathrm{U}}^\top \Sigma^{-1} \widehat{\mathrm{U}} \alpha_i, \qquad\qquad \mathrm{SS}_{\mathrm{res}} = \sum_{ij} \boldsymbol{\varepsilon}_{ij}^\top \Sigma^{-1} \boldsymbol{\varepsilon}_{ij}, \end{aligned} \tag{7}$$

where $\mu = (n_i n_j)^{-1} \sum_{ij} \mu_{ij}$; $\mathrm{SS}_{\mathrm{sce}}$ and $\mathrm{SS}_{\mathrm{ind}}$ are computed similarly using β_j and ϕ_{ij}.

The decomposition of the sum of squares for different number of PCs in $\widehat{\mathrm{U}}$ is illustrated in Fig. 1. An example of the decomposition using 8 PCs for CM4 with the A1B emissions scenario is given in Fig. 2, and Fig. 3 shows the decomposition of NEE into effects that are due to either the GCMs or the emission scenarios.

Results

As seen in Fig. 1 spatial patterns that can be attributed to the GCMs consistently explain more of the variability in NEE fluxes than patterns that are due to the emission scenarios. Fig. 3 clearly illustrates that the contributions from GCMs are both more spatially diverse and larger than the contributions from the scenarios. The largest differences between the GCMs is over the Amazon rain-forest, with lesser deviations for the African rain-forest, northern Eurasia, Alaska, and Canada. The differences over the Amazon are particularly concerning since this is one of the Earth's most productive ecosystems.

This preliminary study illustrates that ecosystem responses simulated under different climate scenarios might have more in common with the GCMs used than with the emissions scenarios. Further study is needed, both to fully quantify these differences and to further investigate why different GCMs give substantially different ecosystems responses.

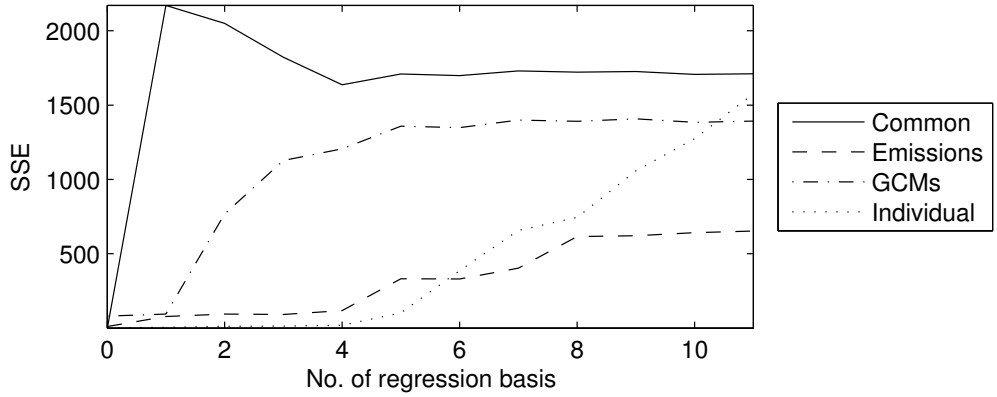

Figure 1: The amount of the SSE explained by each part of the ANOVA, as a function of the number of PCs used in regression basis; zero denotes only an intercept. For reference $\mathrm{SS}_{\mathrm{tot}} = 11'020$ due to the noisy data.

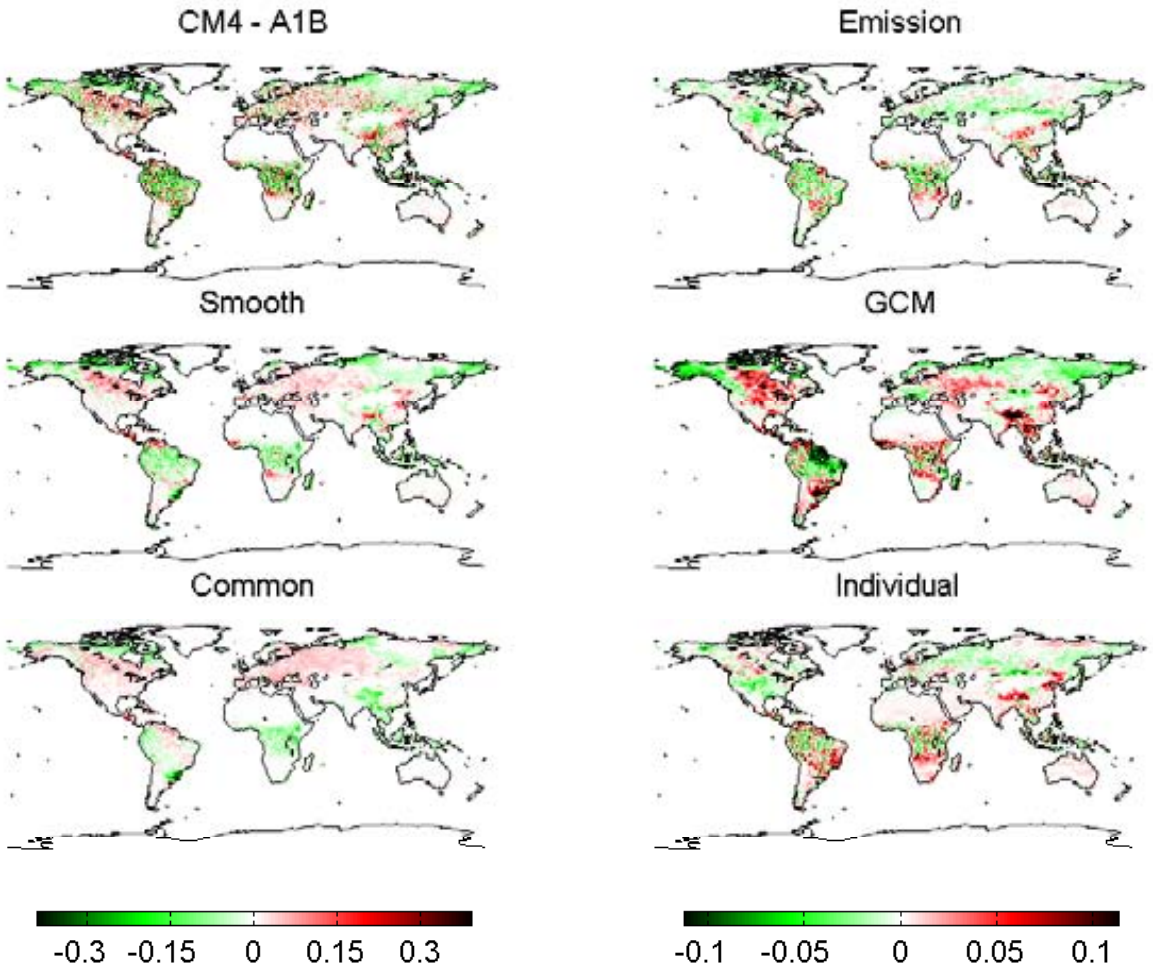

Figure 2: Decomposition of the NEE when LPJ-GUESS is driven by the output from CM4 using the A1B scenario (negative values indicate uptake). From top to bottom on the left, we show the original data, the smooth, and the common contribution (Uc). On the right we have the emission scenario ($\mathrm{U}\alpha_{\mathrm{i}}$), the GCM ($\mathrm{U}\beta_{\mathrm{j}}$), and the individual ($\mathrm{U}\alpha\beta_{\mathrm{ij}}$) contribution. We see that the effect due to the GCM is much larger than the effect due to the scenario. Note that the colour-scale differs between the left and right column.

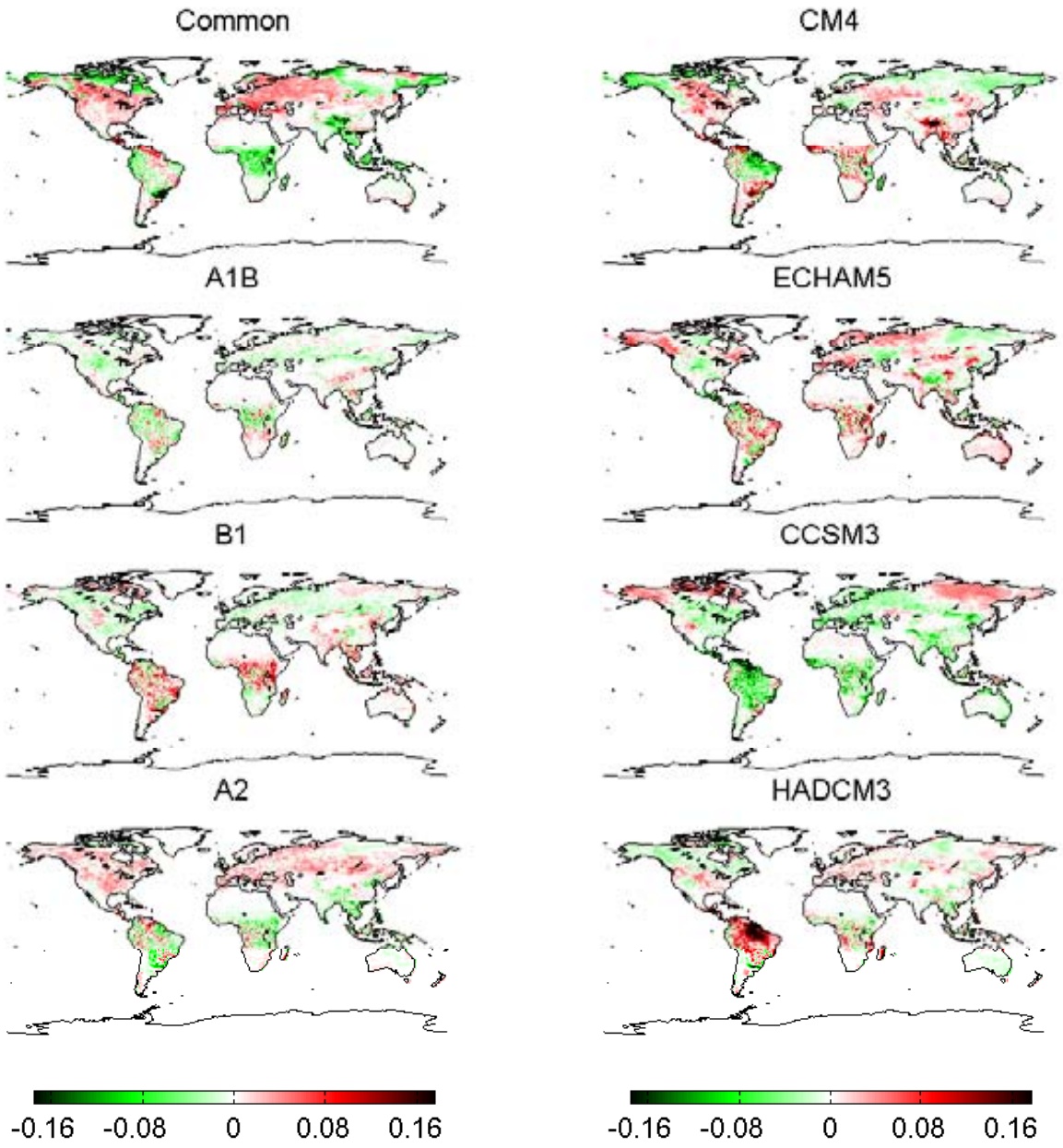

Figure 3: The different contributions to the NEE (negative values indicate uptake) highlighted by the ANOVA. From top to bottom on the left the effect common to all 12 cases, followed by the effects due to the three different emission scenarios is shown. On the right the effect due to the different GMCs is illustrated. The difference between GCMs is much larger than the difference due to emission scenarios.

References (Références)

Aguiar, P. M. & Moura, J. M. (2003). Rank 1 weighted factorization for 3d structure recovery: algorithms and performance analysis. IEEE Trans. Pattern Anal. Machine Intell. 25, 1134–1149.

Boer, G. & Yu, B. (2003). Climate sensitivity and response. Climate Dynamics 20, 415–429.

Collins, W. D., Bitz, C. M., Blackmon, M. L., Bonan, G. B., Bretherton, C. S., Carton, J. A., Chang, P., Doney, S. C., Hack, J. J., Henderson, T. B., Kiehl, J. T., Large, W. G., McKenna, D. S., Santer, B. D. & Smith, R. D. (2006). The community climate system model version 3 (CCSM3). J. Climate 19, 2122–2143.

Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A. & Totterdell, I. J. (2000). Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. Nature 408, 184–187.

Duchon, J. (1976). Splines minimizing rotation invariant seminorms in sobolev spaces. In W. Schempp & K. Zeller, eds., Constructive theory of functions of several variables. Springer-Verlag, pp. 85–100.

Gordon, C., Cooper, C., Senior, C. A., Banks, H., Gregory, J. M., Johns, T. C., Mitchell, J. F. B. & Wood, R. A. (2000). The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Climate Dynamics 16, 147–168.

Kimeldorf, G. & Wahba, G. (1970). A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. Ann. Statist. 41, 495–502.

Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. Roy. Statist. Soc. Ser. B , To appear.

Marti, O., Braconnot, P., Bellier, J., Benshila, R., Bony, S., Brockmann, P., Cadule, P., Caubel, A., Denvil, S., Dufresne, J.-L., Fairhead, L., Filiberti, M.-A., Foujols, M.-A., T. Fichefet, T., Friedlingstein, P., Gosse, H., Grandpeix, J.-Y., F. Hourdin, F., Krinner, G., Lévy, C., Madec, G., Musat, I., de Noblet, N., Polcher, J. & Talandier, C. (2006). The new IPSL climate system model: IPSL-CM4. Tech. Rep. 26, Institut Pierre-Simon Laplace (IPSL), Paris, France.

Mitchell, T. D. & Jones, P. D. (2005). An improved method of constructing a database of monthly climate observations and associated high-resolution grids. Int. J. Climatology 25, 693–712.

Nakicenovic, N. & Swart, R., eds. (2000). Special report on emissions scenarios: A special report of working group III of the IPCC. Cambridge University Press.

Nychka, D. W. (2000). Spatial-process estimates as smoothers. In M. G. A. Schimek, ed., Smoothing and regression: Approaches, computation, and application. Wiley, New York, USA, pp. 393–424.

Quadrelli, R. & Wallace, J. M. (2004). A simplified linear framework for interpreting patterns of northern hemisphere wintertime climate variability. J. of Climate 17, 3728–3744.

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kirchner, I., Kornblueh, L., Manzini, E., Rhodin, A., Schlese, U., Schulzweida, U. & Tompkins, A. (2003). The atmospheric general circulation model ECHAM 5. PART I: Model description. Tech. Rep. ISSN:0937–1060, Max-Planck-Institute for Meteorology, Hamburg, Germany.

Rue, H. & Held, L. (2005). Gaussian Markov random fields; theory and applications, vol. 104 of Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Smith, B., Prentice, I. C. & Sykes, M. T. (2001). Representation of vegetation dynamics in the modelling of terrestrial ecosystems: comparing two contrasting approaches within European climate space. Global Ecology and Biogeography 10, 621–637.

Wahba, G. (1981). Spline interpolation and smoothing on the sphere. SIAM J. Scient. Statist. Comput. 2, 5–16.

Whittle, P. (1954). On stationary processes in the plane. Biometrika 41, 434–449.

Whittle, P. (1963). Stochastic processes in several dimensions. Bull. Internat. Statist. Inst. 40, 974–994.