Efficient treatment of cross-scale interactions in a land-use model

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

d o c t o r r e r u m n a t u r a l i u m
(Dr. rer. nat.)
im Fach Physik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
der Humboldt-Universität zu Berlin

von
Dipl.-Phys. Jan Philipp Dietrich

Präsident der der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Andreas Herrmann

Gutachter:
1. Prof. Dr. Dr. h.c. Jürgen Kurths
2. Prof. Dr. Hermann Held
3. Prof. Dr. Karlheinz Erb

Tag der mündlichen Prüfung: 13. Oktober 2011

This thesis is dedicated to Kerstin
for being at my side all the time.

Acknowledgements

Writing a PhD thesis is a demanding challenge, a maze with pit falls and dead ends. To find a way through it is blessing to have friends and colleagues who are guiding you and are giving a hand. I am glad that many people stand at my side and I want to say thank you!

- My parents Thomas and Gunhild Dietrich. Besides all the other good things they have done to me they gave me the chance to follow my interests and to study.
- My sister Wiebke Dietrich for spending her spare time for wading through all this text searching for mistakes and inconsistencies without any interest in the highlighted topic.
- Kerstin Paul for always showing me that life consists of more than working and studying.
- Michael Winkler for always giving me advice when I needed the unbiased view of an external person.
- Jürgen Kurths for agreeing to become my PhD supervisor and giving me valuable advice.
- Hermann Lotze-Campen and Alexander Popp for supervising my work and giving me full support in everything I was doing.
- Christoph Schmitz for being a wonderful colleague and friend for the wonderful smooth collaboration in all the projects we were handling together.
- Katharina Waha for being such a lovely office roommate suffering all my vocal telephone calls and conversations and for always taking the time to discuss with me my current issues and thoughts.
- Anne Biewald, Benjamin Bodirsky, David Klein, Michael Krause, Christoph Müller, Susanne Rolinski, Isabelle Weindl and all the other friends and colleagues from PIK for always having an open ear for any question or need I had.

Abstract

Due to significant advances in information technology in the last decades complex computer models have become a common tool in various disciplines. One application is the modeling of global land use. A major challenge in modeling is the linking of processes on different scales - such as in land use the local production of agricultural commodities and their global trading. Neglecting these cross-scale interactions leads to significant biases in model projections while a 1:1 representation is computational infeasible. Therefore, a good balance between accuracy and abstraction is essential.

In this thesis I investigate efficient implementations of cross-scale interactions in agricultural land-use models. Based on the global land-use model MAgPIE ("Model of Agricultural Production and its Impact on the Environment") I focus on two dominant aspects: First, the inclusion of spatially explicit data in a global optimization model; second, the proper representation of technological change as a major cross-scale interaction and dominant driver for agricultural land use change.

As a consequence of limitations in complexity of global optimization models the problem arises that spatially explicit, high-resolution data cannot be used directly as model input. Typically, the spatially explicit data is upscaled by using a static upscaling rule which leads to a significant loss of information. As an alternative I discuss the use of clustering methods for upscaling. I provide a general framework including the creation of clusters, the upscaling of inputs, and the downscaling of outputs. My investigations show that the information loss due to upscaling in the upscaled data itself, but also in the model outputs derived with upscaled data decreases significantly compared to the information loss in upscaling with static grids.

Technological change is another important cross-scale interaction. In agriculture technological change means local yield growth induced by supra-regional investments in Research and Development (R&D). Whereas in the past increases in agricultural production were often mainly achieved by expansion of agricultural land, nowadays most increases in total production are outcome of R&D. I present an implementation of this process in MAgPIE including a feedback of land-use intensity on the effectiveness of R&D. To model this feedback I introduce an output-oriented measure for agricultural land-use intensity which takes all sources of intensification into consideration. Based on this measure I show that the effectiveness of investments in R&D decreases with the agricultural land-use intensity. I use this finding to provide an implementation of technological change in MAgPIE.

My findings imply that apart from detailedness especially the implementation has a significant impact on general model quality. Therefore, in model development the framework used for implementation should be emphasized to a greater extent.

Zusammenfassung

Die Fortschritte im IT-Sektor der letzten Jahrzehnte haben komplexe Computermodelle zu einem Standardwerkzeug in vielen wissenschaftlichen Disziplinen werden lassen. Eines der Anwendungsgebiete ist die Modellierung globaler Landnutzung. Ein Hauptzweck von Modellierung ist die Verknüpfung von Prozessen verschiedener Skalen, wie z.B. in der Landnutzung die lokale Produktion landwirtschaftlicher Güter und ihr internationaler Handel. Verzichtet man auf die Verknüpfung dieser Prozesse im Modell, so sind realistische Prognosen in vielen Fällen nahezu ausgeschlossen, bildet man die Realität 1:1 nach, so ist das Modell aufgrund seiner Komplexität nicht mehr lösbar. Kernbedürfnis ist daher eine gute Balance zwischen Genauigkeit und Abstraktion.

In der vorliegenden Arbeit untersuche ich Möglichkeiten, Interaktionen zwischen verschiedenen Skalen in der Modellierung von Landnutzung effizient zu implementieren. Ich fokussiere dabei Prozesse, welche die Dynamik am stärksten bestimmen. Basierend auf dem globalen Landnutungsmodell MAgPIE ("Model of Agricultural Production and its Impact on the Environment") bearbeite ich zwei dominante Prozesse zwischen Skalen: Zum einen die Nutzung hochaufgelöster, räumlich expliziter Daten in einem globalen Modell zur Landnutzungsoptimierung, zum anderen die Modellierung von technologischem Wandel als dominantem Treiber für Landnutzungswandel.

Aus der limitierten Modellkomplexität resultiert das Problem, dass hochaufgelöste, räumlich explizite Daten nicht direkt im Modell genutzt werden können. Meist wird dieses Problem gelöst, indem die Daten nach einem statischen Aggregationsschema hochskaliert werden, einhergehend mit einem starken Informationsverlust. Als Alternative diskutiere ich die Verwendung von Clusteralgorithmen zur Hochskalierung. Ich präsentiere den vollständigen Skalierungsprozess samt der Bestimmung der Cluster, der Aggregation und der nachfolgenden Disaggregation. Meine Untersuchungen zeigen, dass der durch die Hochskalierung verursachte Informationsverlust für die skalierten Daten selbst, aber auch für die Ergebnisse, welche mithilfe der aggregierten Daten simuliert wurden, unter der Verwendung von Clusteralgorithmen signifikant geringer sind als bei der Verwendung statischer Aggregationsvorschriften.

Technologischer Wandel ist eine weitere dominante Interaktion zwischen Skalen in der Landnutzungsmodellierung. In der Landwirtschaft bedeutet technologischer Wandel, dass lokales Ertragswachstum durch überregionale Investitionen in Forschung und Entwicklung (Research & Development - R&D) generiert wird. Während in der Vergangenheit Steigerungen in der Gesamtproduktion der Landwirtschaft zumeist durch Landexpansion erreicht wurden, so geschieht dies heutzutage hauptsächlich durch R&D. Ich präsentiere eine Implementierung dieses Prozesses in MAgPIE mitsamt der Rückkopplung der Landnutzungsintensität auf die Effektivität von R&D. Grundlage dafür ist ein neuentwickeltes Maß für landwirtschaftliche Landnutungsintensität, welches eine umfassende Berücksichtigung aller Einflussgrößen erlaubt. Basierend auf diesem Maß zeige ich, dass die Effektivität von R&D-Investitionen mit steigender Landnutzungsintensität sinkt und stelle die entsprechende Modellimplementierung in MAgPIE vor.

Meine Arbeit zeigt, dass außer dem Detailgrad eines Modells auch die Struktur der verwendeten Implementierungen einen signifikanten Einfluss auf die generelle Qualität der Simulation hat und insgesamt mehr Beachtung in der Modellierung finden sollte.

Contents

1 Preamble

1.1 Motivation

In my diploma thesis my former supervisor Prof. Bernd Blasius gave me the wonderful opportunity to do research on a generalization of an extremely powerful data analysis method called Phase Space Reconstruction [Dietrich, 2008]. I was starting my work with huge optimism and motivation driven by the power of this method. But when I had finished the methodology part of my thesis and was starting to search for applications I noticed that despite the fact that this method was powerful I had serious problems to find examples where I could apply it to.

At this point I realized that there often seems to exist a huge gap between theoretical analysis methods and applications. Simplified and a bit overstated I would describe it as follows: On one side there is a group of scientists (group A) who is inventing and constructing clever data analysis methods which allow to receive information hidden in the data. On the other side there are scientists (group B) who work directly at explicit problems that they have to analyze. However, the interaction between both groups is lacking. Group A has often problems to find proper applications for its methods whereas group B does not know anything about the methods developed in group A. So the first group is inventing powerful tools which are used only sporadic and the second group is doing semi-optimal analysis of explicit problems because of a lacking expertise in the usage of tools delevoped by the first group.

The experience to see that there is a powerful tool which offers excellent opportunities, but which is probably never used because scientists who would benefit of this method probably will never hear something about it, was very frustrating to me. So I decided to search for options to overcome that gap. The solution I tried in my PhD time was just to shift from group A to group B. I thought it could make sense for me as a theoretical data analysis physicist to apply for a job in a group of applied scientists. My aim was to use my knowledge about general data analysis methods to apply it for that special problem my group was working on.

I am glad that I found a wonderful group of scientists which was willing to hire me even though I was not fitting to the announced vacancy and which offered me the opportunity to do that experiment. What I experienced was again that it is not as easy as expected. Actually I found out that many of that analysis methods I learned about have significant limitations. Probably the most problematic one was that most methods need a huge amount of data which is not a problem applied on idealized systems, but in many cases for real applications the accessible data is often strongly limited and of low quality (this especially holds true when working with economic data).

Another problem are the scientific expectations: When applying for a job in a group

doing research on tools and methods it is expected that one will invent new methods. When applying in a group that is doing research on a explicit problem it is expected that one will produce results concerning this explicit problem. But working in an applied group and doing work on implementing theoretical methods to increase the general quality of outputs this kind of research is hard to sell. Because the quality of scientific work is measured in papers and citations the benefit of a general internal quality improvement is strongly limited. Hence most scientists in applied science groups have to focus more on the direct output than on general quality improvements. For me as a theoretical physicist in applied science it was a kind of balancing act: just switching the side does not erase the gap. Instead there is often the risk to fall into the gap. Instead of being part of both groups it can also happen that one does not belong to any of them. Nevertheless I recommend everybody to switch sides because only in this way it is possible to get behind the problems that are causing that gap and to understand what both sides can deliver and what both sides need.

1.2 Thesis in context of Physics

This thesis deals with the development of an economic, agricultural land-use model. It touches many disciplines like agriculture, economics, geography, ecology, hydrology, and climatology, so one might ask: Where is the physics in it? What does it make to be a PhD thesis in physics? The answer lies in the used methodology and applied perspective that I used to analyze and tackle the problems. I used approaches that are common for physicists but widely unknown or unusual in the field of land-use modeling. Typical examples for these kind of approaches are information theory (the focus on information conservation in an abstract, content-unrelated sense - see Chapter 3), mapping of processes to scales (determination of scales on which an examined process plays a dominant role - see Chapter 2.1), and the use of well-defined measures (see Chapter 4).

While I had to realize quite fast that many tools used in physics are not applicable on research questions dealing with the econonmics of agricultural land-use, I also made rather quickly the experience that several general concepts of physics are quite useful. Especially the approach to simplify and generalize process descriptions is a so far mostly unused concept in agricultural land-use modeling and a powerful counterpart to the prevailing “more detail = more accuracy”-model. In general, focusing on the underlying model processes from a model-theoretical point of view of a physicist while the rest of the team is focusing on its contents emerged to be a quite complementary approach. Not the topics make this thesis belonging primarily to Physics. It is the used point of view and the applied methodologies and concepts which reason its relation.

One major difference between agricultural land-use modeling and other research topics which are more closely connected to natural sciences is the degree of underlying uncertainties. In land-use modeling one has to face a situation which is typical for research closely connected to social sciences: Uncertainties are extremely high so that most findings can only be seen as a best guess rather than a robust finding.

Being unaware of its origin one might interpret it as a consequence of bumbling re-

search. However, there are several serious reasons which are explaining this fact rationally:

First of all, the model results and applied assumptions typically cannot be verified with experiments. Most outputs describe supra-regional dynamics which can only be verified by historic observations. This is equal to an experiment under inaccurately defined boundary conditions which can only be performed once. Hence, this reference can only deliver quite limited insights in terms of verification of the model results. There are also real experiments taking place in agriculture, but the measured outputs, as for instance yields, are typically significantly higher than yields observed in practice because of ideal boundary conditions. So, their results have also a quite limited usefulness for agricultural models.

Second, the agricultural model community has to face, as many other communities related to social sciences, the situation of poor data availability and poor data quality. Most parameters relevant for the model dynamic cannot be measured at all and therefore, they are only results of other models. And even the few numbers that are not simulation results, as for instance crop-specific production information on country level as supplied by FAO [FAOSTAT, 2009], are often only indirect measurements or expert guesses. Since this data typically bases on national statistics their quality varies from country to country. Some countries deliver quite detailed and reliable data while data from other countries has to be assessed by experts as these countries deliver no data at all. So even data from the best available sources in agriculture already contains a high degree of uncertainty on an aggregation level which is much coarser than the level the information is actually required for.

Third, as any research connected to social sciences humans are part of the concept. In general, this does not inhibit the finding of robust and general dynamics in the system, but complicates it because any human being has its own will which is strongly influenced by the current social environment one is living in. Furthermore, the publication of model results can cause a feedback to the observed system which is then changing the dynamics of it. However, this point can be disregarded in most cases at the current stage since it is typically outshined by the other major sources of uncertainty in agricultural research mentioned before.

In awareness of all these problems and uncertainties in agricultural modeling one might ask if it makes sense to do further research on it since it seems to be impossible to achieve robust insights. Unfortunately, this is not an option due to an urgent need by decision makers. Unlike many other research fields in natural sciences, insights concerning the agricultural sector are time-critical since necessary decisions have to and will be taken now and cannot be delayed into the future. From this point of view a best guess, as it can be supplied by agricultural land-use models, is worse than a robust finding, but even better than no information concerning the agricultural sector at all.

Summarized, it is extremely important to have these issues in mind when dealing with agricultural land-use models. It is the reasoning for some implementations that seem to be odd, but it is also a warning of overrating statements derived from agricultural models.

2 Introduction

Abstract

In this chapter I give a brief review about the topics relevant for this thesis. Before presenting the general outline I introduce the basic concepts and applied tools: After explaining what the term "scale" means and what "land-use modeling" stands for I present some basics about the mainly used agricultural land-use model "Model of Agricultural Production and its Impact on the Environment" (MAgPIE), the global vegetation model "Lund-Potsdam-Jena with managed Land" (LPJmL), which is delivering several data sets used for the following analysis, and the "General Algebraic Modeling System" (GAMS), the programming language in which the MAgPIE core is written. Finally, I explain in the outline how these different topics are interrelated and what corresponding research questions I tackle in this thesis.

2.1 The scale concept

An important step in the analysis of a process is its structuring into sub-process or sub-objects. For the separation it is necessary to have some kind of classification scheme. One very helpful approach is to classify processes or objects based on the scale they are related to. According to Turner et al. [2001, p. 27] "scale refers to the spatial or temporal dimension of an object or process". In the case of objects it is related to their size or life-time, in the case of processes it is related to a characteristic time span or spatial extend, for instance the period time or period length of a periodic process.

To describe a band of several scales, for instance to characterize the range of scales covered by a model, two further terms are used: grain and extend [Turner et al., 2001, Krüger, 2007]. Grain is the smallest unit (temporal or spatial) of a data set, model, or an observation, for instance the grid size of an spatial explicit data set. It marks the lower end of the covered scales. Extend describes the total spatial or temporal coverage and is the upper scale limit.

Another important term in this context is resolution, which stands for the precision or detailedness of an measurement [Turner et al., 2001]. Often it is used interchangeably with grain since grain is an indicator for the resolution of a measurement [Gibson et al., 2000, Turner et al., 2001]. However, in some situations this relation does not hold. An example are disturbances in a measurement, so that differences between adjacent cells are outshined by the noise. In that case a further decrease in grain size does not deliver additional detail and does not lead to an increased resolution. Grain size is always related to the physical characteristics of a data set (size of a single data point), whereas resolution refers to the quality of a data set (the detailedness in which the original system

is reproduced by the data set). This differentiation is especially relevant for the work presented in chapter 3 of this thesis.

To understand problems, that come along with the scale issue, it is also helpful to distinguish between different kinds of scales. In literature typically three kinds of scales are distinguished: the process scale, the observation scale and the modeling (or working) scale [Blöschl and Sivapalan, 1995, Krüger, 2007]. The first one is the scale a process can be identified with, based on its characteristics. This scale is predetermined by the process. The second one is the scale on which observations are performed. This scale can be chosen by the researcher and should in the ideal case agree with the process scale. The last scale is the modeling scale on which processes are described by the researcher. This scale is also customizable and is typically adjusted to process and observation scale, but also based on the underlying research questions. Harmonizing these scales is one big challenge modelers have to face (Chapter 3).

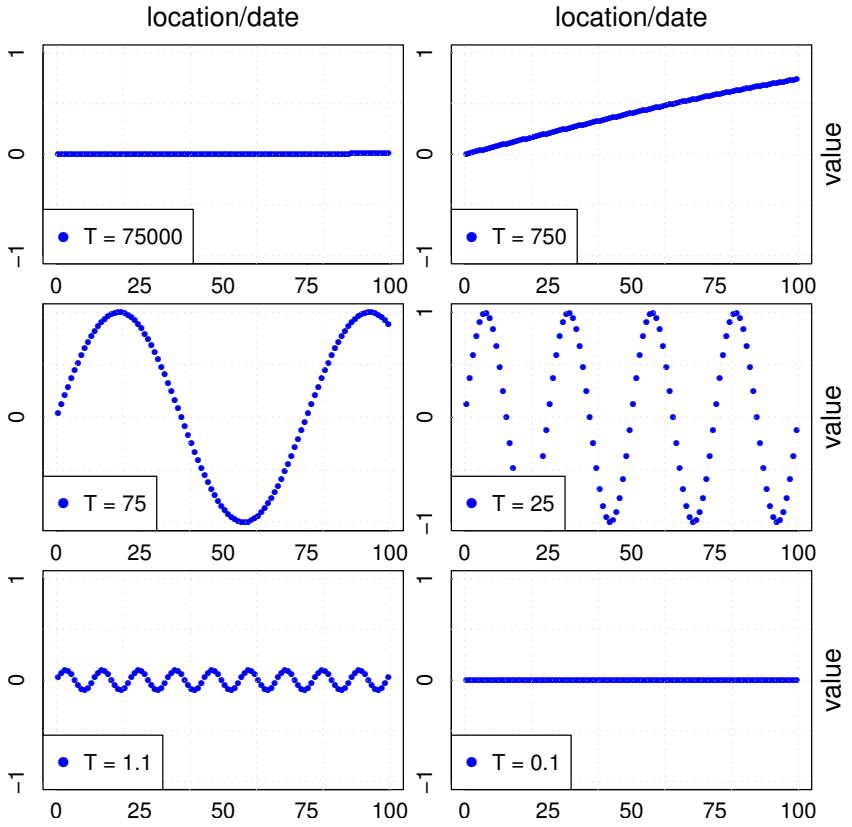

Figure 2.1: Harmonic oscillators with different period times/lengths T and oscillations from -1 to 1 shown for a system with grain size 1 and extend 100.

In the absence of nonlinear effects it is possible to detach processes acting on different scales from each other completely and to analyze them separately. This is possible because processes relevant at one scale become irrelevant for other scales. Harmonic oscillators are a good example to explain this behavior.

Figure 2.1 shows harmonic oscillators with different periods observed in a system with grain size 1 and extend 100 (either temporal or spatial units). The marginal distance between grain size and extend assures, that only a small band of scales is observed. Starting with a very long period of $T = 75000$ the process value remains nearly constant. The applied observation scale is too small to show any dynamics. For $T = 750$ an increase of the value becomes visible. However, this is only a part of the full dynamic of the harmonic oscillator, the observation scale is still much lower than the process scale. In contrast, the harmonic oscillators with $T = 75$ and $T = 25$ are fully observable at the applied observation scale, since at least one full period is performed within the extent. The oscillator with $T = 1.1$ shows, what happens if the period is further decreased. In that case the extend of the observation does no play any role, since it is big enough to capture a full period. Instead the grain size, which is representing the lower end of observed scales, comes into play. An observation at point x is the mean of the harmonic oscillators values between $x - 0.5$ and $x + 0.5$. Having a periodicity close to the grain size leads to the situation, that parts of the model dynamic are averaged out. The model dynamic is falsified which is reflected by a reduced amplitude. Finally, a further decrease to $T = 0.1$ leads to the same situation already observed for too long periods: no dynamic is visible anymore. However, in this case the behavior is caused by dynamics with a too high periodicity, which leads to a complete clearing of the dynamic at the observed scale.

The example shows a system in which only processes become relevant, which have a process scale close to the chosen observation scale. For oscillators with process scales much bigger oder smaller compared to the observation scale no dynamic is observable. Thus, these processes can be neglected for further investigations. However, not all processes can be separated only because they are occurring on different scales. Due to nonlinearities processes of one scale can influence processes and results on another scale. For instance, a burning match can cause a whole forest fire. These interactions between scales (linear and nonlinear) are covered by the term cross-scale interactions. Whereas linear cross-scale interactions are quite simple to handle, as the problems are completely separable, nonlinear cross-scale interactions can lead to serious problems and model biases.

Cross-scale interactions play an important role in global change research [Wessman, 1992, Cash and Moser, 2000, Harvey, 2000] for several reasons. First, the integration of models and data from different disciplines, such as Physics, Biology, Geography or Economy, is typically connected to the issue of different spatial and temporal scales [Wessman, 1992]. Second, because of nonlinearities a proper treatment of cross-scale interactions is often a requirement for accurate simulations [Cash and Moser, 2000, Harvey, 2000]. Third, the interactions itself are of great interest to understand the dynamics and to be able to assess the impact of policies at different scales [Cash and Moser, 2000, Dirnböck et al., 2008].

2.2 Land-use modeling

Land-use models are models which address the allocation of land to specific land-use types. They are used to make projections of future land-use patterns and to simulate land-use change under various scenarios. Today the most prominent land-use type is agricultural land (38% of total land area in 2008) followed by forest area (31%) [FAOSTAT, 2011]. Approximately 1/3 of agricultural land is arable land used for crop production, while the rest is pastureland used for livestock production. Based on the huge area under agricultural production and the strong interrelation of population driven demands and agricultural production the sub-class of agricultural land-use models evolved.

Agricultural land-use models are quite similar to general land-use models. The only difference is, that the agricultural land is modeled in more detail compared to other land-types, splitting cropland in several sub-types with different crops. Furthermore, these models typically also provide a more detailed representation of the demand side for agricultural commodities, as they are acting as drivers for agricultural land cover change. Basic questions agricultural land-use models try to answer are: Can the agricultural sector meet future demands under scenario X? What land cover change can we expect under scenario Y? What does that mean in terms of deforestation? How will food prices react? Which role will biofuels play in the future? How will climate change affect agricultural production and land cover change?

2.3 The MAgPIE model

The model this thesis deals with, is called "Model of Agricultural Production and its Impact on the Environment" (MAgPIE) . It is a nonlinear, recursive dynamic, agricultural land-use model, that links regional economic information with grid-based biophysical constraints [Lotze-Campen et al., 2008, 2010, Popp et al., 2010]. In each time step a term describing the total costs of production is minimized. The results of a previous time step are used as inputs for the current time step and supplemented by a set of time depending and time independent parameters.

All model calculations are taking place at one of three spatial scales: A global scale representing global markets, a regional scale of 10 world regions representing specific economic development, demands, technology levels and trade, and a local scale with a grain size of up to $0.5^\circ \times 0.5^\circ$ representing farming decisions based on spatially varying production parameters, such as potential yields or water availability. Figure 2.2 shows the regional coverage of the 10 world regions, the corresponding mapping between world regions and countries is attached in the appendix (Table 2).

Currently, there exist several MAgPIE derivatives with different focuses, such as global trade, livestock production or emission policies. All of them base on a main MAgPIE version (trunk), which is described in the following. Hence, most statements issued in the following apply also for these derivatives. Furthermore, the model improvements, described in the following chapters 3, 4, and 5 are also part of all currently developed derivatives.

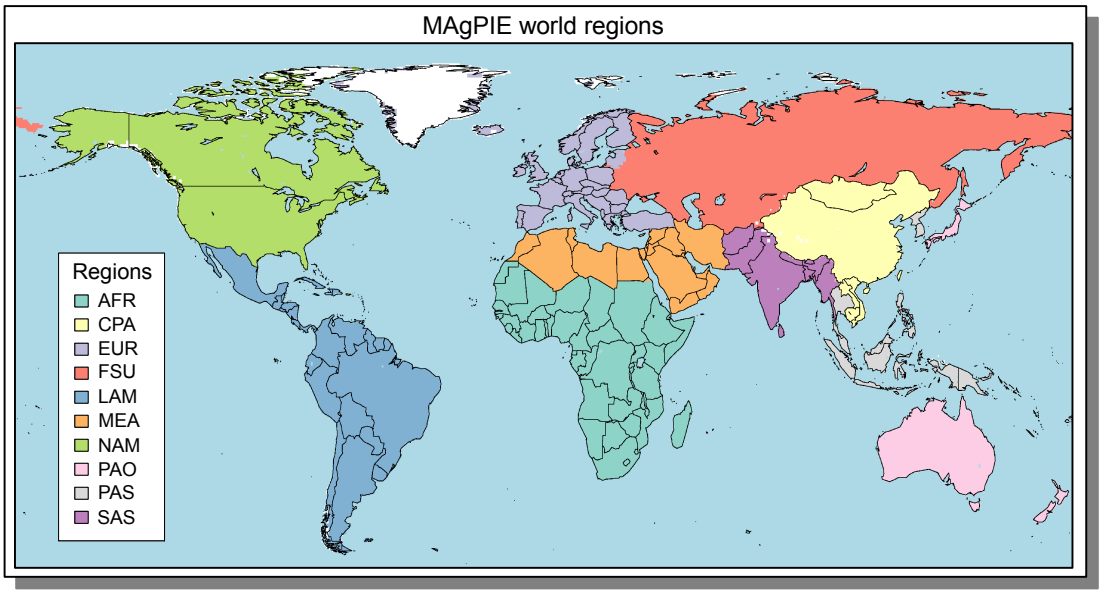

Figure 2.2: MAgPIE world regions: AFR = Sub-Sahara Africa, CPA = Centrally Planned Asia (incl. China), EUR = Europe (incl. Turkey), FSU = Former Soviet Union, LAM = Latin America, MEA = Middle East and North Africa, NAM = North America, PAO = Pacific OECD (Australia, Japan and New Zealand), PAS = Pacific Asia, SAS = South Asia (incl. India)

Mathematically a simulation run of MAgPIE in the simulation period T can be described as a set of solutions X_t of a time depending minimization problem (Equation 2.1).

$$X_t = \{x_t \mid \forall y \in \Omega \wedge t \in T : g_t(x_t) \leq g_t(y)\} \subseteq \Omega \tag{2.1}$$

For every timestep $t \in T$ the element x_t minimizes the function value of the goal function $g_t(x_t)$, where the goal function at time step t depends on the solutions of the previous time steps $x_{(t-1)}, ..., x_1$ and a set of time depending parameters P_t (Equation 2.2).

$$g_t(x_t) = g(t, x_t, x_{(t-1)}, ..., x_1, P_t) \tag{2.2}$$

2.3.1 Sets

The dimension of the domain Ω, on which for each timestep the minimization problem is defined, and Ω_T, on which the full system is defined, depend on the following sets:

- $T = \{\text{time steps } t\}$: Time - t stands for the current time step, $t-1$ for the previous time step and so on. The first simulated time step is $t = 1$, which is currently the

year 1995 with an increment of 10 years per time step. The latest time step, that is currently simulated, is 2145.

- I = {world regions i}: Economic world regions in MAgPIE. Currently MAgPIE contains 10 world regions (AFR = Sub-Sahara Africa, CPA = Centrally Planned Asia (incl. China), EUR = Europe (incl. Turkey), FSU = Former Soviet Union, LAM = Latin America, MEA = Middle East and North Africa, NAM = North America, PAO = Pacific OECD (Australia, Japan and New Zealand), PAS = Pacific Asia, SAS = South Asia (incl. India) - Figure 2.2).

- J = {spatial clusters j} : Highest disaggregation level used in the optimization. The number of clusters can be chosen by the user. Typically simulation runs are performed with 100 - 4000 clusters, depending on models run purpose (quick scenario check, elaborated model run for a publication,...) and complexity of the MAgPIE derivative. Before calculations, spatial-explicit input data with a grain size of $0.5^\circ \times 0.5^\circ$ is upscaled to these clusters. After optimization, clusters are brought back to the input grain size by downscaling. The full process of up- and downscaling and the generation of clusters was part of this thesis and is described in chapter 3.

- K = {simulated products k} : Union of vegetal products V and livestock products L ($K = V \cup L$).

- L = {simulated livestock products l}: Products simulated within the livestock sector of MAgPIE (≈ 5, varies between different model versions).

- V = {vegetal products v}: Products simulated within the crop sector of MAgPIE (≈ 20, varies between different model versions).

- W = {water supply types w}: Currently two types are implemented: rainfed 'rf' and irrigation 'ir'

- C = {crop rotation groups c}: Groups of crops, which produce similar effects in terms of crop rotation (≈ 20, varies between different model versions).

The combined model variable x_t consists of three sub-variables x_t^{area}, x_t^{prod} and x_t^{tc} representing different agricultural and economic contents.

$$x_t = \left(x_t^{area} \in \Omega^{area}, x_t^{prod} \in \Omega^{prod}, x_t^{tc} \in \Omega^{tc}\right) \in \Omega, \tag{2.3}$$

The respective domains can be identified as the following vector spaces:

$$\Omega^{area} = \mathbb{R}^{|J|} \times \mathbb{R}^{|V|} \times \mathbb{R}^{|W|} \tag{2.4}$$

$$\Omega^{prod} = \mathbb{R}^{|J|} \times \mathbb{R}^{|L|} \tag{2.5}$$

$$\Omega^{tc} = \mathbb{R}^{|I|} \tag{2.6}$$

Based on this the solution space for each timestep Ω has the dimension $dim\Omega = |J| \cdot |V| \cdot |W| + |J| \cdot |L| + |I|$. Furthermore, the total solution space of a full MAgPIE run $\Omega_T = \Omega \times T$ has the dimension $dim\Omega_T = |T| \cdot dim\Omega = |T| \cdot (|J| \cdot |V| \cdot |W| + |J| \cdot |L| + |I|)$.

In the following, variables and parameters are provided with subscripts to indicate the dimension of the respective subdomains. Subscripts written in quotes are single elements of a set. The order of subscripts in the variable, parameter and function definitions does not change. The names of variables and parameters are written as superscript.

2.3.2 Variables

Since MAgPIE is a recursive dynamic optimization model, all variables refer to a certain time step $t \in T$. In each optimization step, only the variables belonging to the current time step are free variables. For all previous time steps, values were fixed in earlier optimization steps. As I have shown above, we currently distinguish three variables x_t^{area}, x_t^{prod} and x_t^{tc} that can be described as follows:

- $x_{t,j,v,w}^{area}$: The total area of each vegetal production activity v for each water supply type w, each cluster j and each time step t. This variable describes the spatial and temporal land-use allocation dynamics. [10^6 ha]

- $x_{t,j,l}^{prod}$: The total production of each livestock product l, for each cluster j at each time step t. It contains spatial and temporal livestock allocation dynamics [10^6 ton dry matter]

- $x_{t,i}^{tc}$: The amount of yield growth for each time step t in each region i triggered by investments in infrastructure, research and development. It is the counterpart of land expansion. A more detailed discussion of it can be found in chapter 5 based on concepts developed in chapter 4. [1]

2.3.3 Parameters

Besides variables, the model is fed with a set of parameters P_t. These parameters are computed exogenously and are in contrast to variables of previous time steps fully independent of any simulation output. Although most parameters are time independent, there exist also some parameters which are time dependent.

- $p_{t,j,v,w}^{yield}$: Yield potentials for each time step t, cluster j, crop v and water supply w type taking only natural variations into account and excluding changes due to technological change. Data is provided by the global dynamic vegetation model LPJmL (Section 2.4). This parameter is time dependent, since LPJmL can deliver yield potential projections based on various climate scenarios, which cause changes in natural yield potentials over time. [ton/ha]

- $p_{t,i,k}^{dem}$: Regional food and material demand in each time step t for each product k. This data as well as the data for the following parameters base on country-specific

FAO data [FAOSTAT, 2009], which was processed and upscaled to MAgPIE regions and production types. [10^6 ton]

- $p^{fshr}_{i,l,k}$: Feed share describing the regional share of each product k of the total feed production for livestock product l and corresponding transformation from GJ feed in ton dry matter. [ton/GJ]
- $p^{feed}_{i,l}$: Feed requirements for each livestock product l in each region i. [GJ/ton]
- $p^{byprod}_{i,k,l}$: Feed energy delivered by the byproducts of k that are available as feedstock for the livestock product l [GJ/ton]
- $p^{frv}_{i,v}$: Area related factor requirements for each crop v and each region i based on the technological development situation in the initial time step. This parameter as well as the following ones base on outputs from the GTAP model [Narayanan and Walmsley, 2008] and FAO data [FAOSTAT, 2009]. [US$/ha]
- $p^{frl}_{i,l}$: Production related factor requirements for livestock products for each livestock type l and each region i [US$/ton]
- p^{lcc}_{i}: Area related land conversion costs for each region i. This parameter describes the costs which arise, when non-agricultural land is converted to land equipped for agricultural production. [US$/ha]
- p^{tcc}: Technological change costs factor containing an interest correction, an expected lifetime factor and a general cost factor (Chapter 5). [US$/ha]
- $p^{\tau 1}_{i,v}$: τ-Factor representing the agricultural land use intensity in the first simulation time step for each crop v in each region i (based on the research presented in chapter 4).[1]
- p^{cxp}: Elasticity between τ-Factor and investment-yield ratio (Chapter 5) [1]
- $p^{seed}_{i,v}$: Share of production that is used as seed for the next period calculated for each crop v in each region i [1]
- $p^{xs}_{t,i,k}$: Regional excess supply for each product k and each time step t describing the amount produced for export [10^6 ton]
- $p^{sf}_{i,k}$: Regional self sufficiencies for each product k [1]
- p^{tb}: Trade balance reduction factor. It describes the share of the total trade volume, which is distributed according to a fixed trading scheme based on the trade patterns observed in 1995. Increases in the trade volume caused by increases in global demand are distributed proportionally between all exporting regions. Relaxing the trade balance constraint allows the optimizer to deviate from this trading scheme to find a more cost saving trading pattern. [1]

- p_j^{land}: Total amount of land available for crop production in each cluster j. Land in this category does not belong to a competing land use class, such as forest, pasture or urban land and its environmental conditions do not inhibit the use for agricultural production. [10^6 ha]
- $p_j^{ir.land}$: Total amount of land equipped for irrigation in each cluster j. These are areas, which are equipped with an irrigation system. [10^6 ha]
- $p_{j,k}^{watreq}$: Cluster-specific water requirements for each product k. For livestock products this is the total amount of water required. In the case of vegetal products it is the amount of water, that has to be added to the precipitation water, to achieve optimal plant irrigation. [$m^3/ton/a$]
- p_j^{water}: Amount of water available for production in each cluster j (for livestock production and irrigation). [m^3/a]
- p_c^{rmax}: Maximum share of crop groups in relation to total agricultural area [1]
- p_c^{rmin}: Minimum share of crop groups in relation to total agricultural area [1]

[all ton units in dry matter]

2.3.4 Sub-functions

To lighten the general model structure, some model components which appear more than once in the model description and depend on the variables of the current time step t are arranged as functions:

$$f_{t,i}^{growth}(x_t) = \prod_{\tau=1}^{t} (1 + x_{\tau,i}^{tc}) \tag{2.7}$$

$$f_{t,i,k}^{prod}(x_t) = \sum_{j_i} \begin{cases} x_{t,j,k}^{prod} & : k \in L \\ \sum_w x_{t,j,k,w}^{area} p_{t,j,k,w}^{yield} f_{t,i}^{growth}(x_t) & : k \in V \end{cases} \tag{2.8}$$

$$f_{t,i,k}^{dem}(x_t) = p_{t,i,k}^{dem} + \sum_l p_{i,l,k}^{fshr} \left(p_{i,l}^{feed} f_{t,i,l}^{prod}(x_t) - \sum_\kappa p_{i,\kappa,l}^{byprod} f_{t,i,\kappa}^{prod}(x_t) \right). \tag{2.9}$$

- $f_{t,i}^{growth}$: Growth function describing the aggregated yield amplification due to technological change (Chapter 5) compared to the level in the starting year 1995 for each year t and region i. [1]
- $f_{t,i,k}^{prod}$: Function representing the total regional production of a product k in region i at timestep t. In the case of vegetal products, it is derived by multiplying the current yield level with the total area used to produce this product. In the case of livestock products, it is represented by the related production variable. [10^6 ton dry matter]

- $f^{dem}_{t,i,k}$: Function defining the demand for product k in region i at timestep t. It consists of an exogenous demand for food and materials $p^{dem}_{t,i,k}$ and an endogenous demand for feed, which is calculated as the feed demand generated by the livestock production minus the feed supply gained through byproducts. [10^6 ton dry matter]

2.3.5 Goal function

$$g_t(x_t) = g(t, x_t, x_{(t-1)}, ..., x_1, P_t) \tag{2.10}$$

The goal function describes the value that is minimized during the optimization in each time step. It is time-dependent, meaning that it looks different for each time step depending on the solutions of the previous time steps and the time depending parameters. The goal function $g_t(x_t)$ is defined as follows:

$$\begin{aligned} g_t(x_t) = & \sum_{i,v} \left(p^{frv}_{i,v} f^{growth}_{t,i}(x_t) \sum_{j_i,w} x^{area}_{t,j,v,w} \right) \\ & + \sum_{i,l} \left(p^{frl}_{i,l} f^{prod}_{t,i,l}(x_t) \right) \\ & + \sum_{i} \left(p^{lcc}_{i} \sum_{j_i,v,w} \left(x^{area}_{t,j,v,w} - x^{area}_{t-1,j,v,w} \right) \right) \\ & + p^{tcc} \sum_{i} \left(x^{tc}_{t,i} \left(\frac{1}{|V|} \sum_{v} p^{\tau 1}_{i,v} f^{growth}_{t,i}(x_t) \right)^{p^{cxp}} \sum_{j_i,v,w} x^{area}_{t-1,j,v,w} \right). \end{aligned} \tag{2.11}$$

The function describes the total costs of agricultural production. The total costs can be split in four terms: First, the area depending factor costs of vegetal production, which increase with the yield gain due to technological development (Chapter 5). Second, the factor costs of livestock production depending on the production output. Third, the land conversion costs which arise when non-agricultural land is cleared and prepared for agricultural production. Fourth, the costs, which arise by investing in technological development to increase yields by new inventions and improvements in management strategies (Chapter 5). The technological change costs are proportional to the total cropland area of a region and increase disproportionate with the yield growth bought in the current time step and the agricultural land-use intensity τ (Chapter 4).

2.3.6 Constraints

Constraints are used to describe the boundary conditions, under which the goal function is minimized.

Global demand constraint (for each activity k)

$$\sum_i \frac{f^{prod}_{t,i,k}(x_t)}{1+p^{seed}_{i,k}} \geq \sum_i f^{dem}_{t,i,k}(x_t) \tag{2.12}$$

This constraint describes the global demand for agricultural commodities: The total production of a commodity k adjusted by the seed share required for the next production iteration has to meet the demand for this product.

Trade balance (for each region i and product k)

$$\frac{f^{prod}_{t,i,k}(x_t)}{1+p^{seed}_{i,k}} \geq p^{tb} \begin{cases} f^{dem}_{t,i,k}(x_t) + p^{xs}_{t,i,k} & : p^{sf}_{i,k} \geq 1 \\ f^{dem}_{t,i,k}(x_t) p^{sf}_{i,k} & : p^{sf}_{i,k} < 1 \end{cases} \tag{2.13}$$

The trade balance constraint is similar to the global demand constraint, except that it acts on a regional level. In the case of an exporting region (self sufficiency for the product k is greater than 1), the production has to meet the domestic demand supplemented by the demand caused due to export. In the case of importing regions (self sufficiency less than 1), the domestic demand is multiplied with the self sufficiency to describe the amount which has to be produced by the region itself. In both cases the demand is multiplied with a so called "trade balance reduction factor". This factor is always less or equal 1 and is used to relax the trade balance constraints depending on the particular trade scenario, that is run.

Land constraints (for each cluster j)

$$\sum_{v,w} x^{area}_{t,j,v,w} \leq p^{land}_j \tag{2.14}$$

$$\sum_{v} x^{area}_{t,j,v,'ir'} \leq p^{ir.land}_j \tag{2.15}$$

The land constraints guarantee, that no more land is used for production than available. The first land constraint (Equation 2.14) ensures the land availability for agricultural production in general. The second one (Equation 2.15) secures, that irrigated crop production is restricted to areas that are equipped for irrigation.

Water constraint (for each cluster j)

$$\sum_v x^{area}_{t,j,v,'ir'} p^{yield}_{t,j,v,'ir'} f^{growth}_{t,i(j)}(x_t) p^{watreq}_{j,v} + \sum_l x^{prod}_{t,j,l} p^{watreq}_{j,l} \leq p^{water}_j \tag{2.16}$$

In MAgPIE, the production of animal commodities as well as vegetal goods produced with irrigation requires water. The required amount of water is proportional to the

production volume. The whole cluster-specific water demand must be less or equal to the water available for production in this cluster.

Rotational constraints (for each crop group c, cluster j and irrigation type w)

$$\sum_{v_c} x^{area}_{t,j,v,w} \leq p^{rmax}_c \sum_{v} x^{area}_{t,j,v,w} \tag{2.17}$$

$$\sum_{v_c} x^{area}_{t,j,v,w} \geq p^{rmin}_c \sum_{v} x^{area}_{t,j,v,w} \tag{2.18}$$

The rotational constraints are used to describe crop rotations, but also other aspects such as cultural preferences or efforts towards autonomic food production systems. This is achieved by defining for each vegetal product a maximum and minimum share relative to total area under production in a cluster. While crop rotation structures are exclusively described with the maximum share constraint, cultural preferences and autonomy efforts are basically described with the minimum constraint.

2.4 The LPJmL model

The "Lund-Potsdam-Jena with managed Land" (LPJmL) model is a process-based, global vegetation model, which simulates natural vegetation with 9 plant functional types (PFTs) [Sitch et al., 2003] as well as agricultural vegetation with currently 16 crop functional types (CFTs) [Bondeau et al., 2007]. The model has a grain size of $0.5° \times 0.5°$ and works on a daily time step basis. An explicitly modeled carbon cycle with its relevant processes such as photosynthesis, carbon allocation and respiration and the implemented water cycle (runoff, discharge, interception, evaporation, transpiration) enables LPJmL to estimate feedbacks of changes in atmospheric CO_2 concentrations, precipitation and temperature changes [Gerten et al., 2004, 2007, Rost et al., 2008]. The process-based implementation of seasonal phenology (sowing and harvest dates) of CFTs allows for adaptation of crop varieties and growing periods to climate change [Bondeau et al., 2007, Waha et al., 2011].

LPJmL provides MAgPIE cellular data ($0.5° \times 0.5°$, yearly) on current and future yield levels (for rainfed production and production with irrigation, crop-specific), crop-specific water requirements for irrigation and cellular water availability based on simulated discharge. The yield data is derived by using an artificial land use pattern giving each CFT in each cell a non-vanishing land use share. Comparing the water use under activated and deactivated water stress (limitation of water supply) gives the water demand needed for irrigation. Water availability is simulated running a scenario with natural vegetation only.

The data is preprocessed according to the MAgPIE format standards (10 year averages, cells partitioned based on MAgPIE regions) and then upscaled using the methodology described in chapter 3. The data is calculated for the time period from 1995 to 2095 based on various climate scenarios. However, in MAgPIE it is also possible to run simulations

assuming static yield levels using only the LPJmL values calculated for 1995.

2.5 General Algebraic Modeling System (GAMS)

Whereas LPJmL is written in C, MAgPIE is written in GAMS ("General Algebraic Modeling System"), a special language designed to express optimization problems [Brook et al., 1988, McCarl et al., 2008]. It is mostly used in economics and its typical applications are models based on cost minimization or profit maximization. In contrast to classical programming languages the code is not executed step by step. Instead one has to define the structure of the model with parameters, constants, variables, constraints that have to be fulfilled and a goal function that has to be minimized or maximized. Executing GAMS this model description is sent to an external solver which compiles the problem and solves the constraints. Consequence of this setup is a strict separation of the numerical part done by the solver and the general model description supplied by the user. On one hand this allows the modeler to use highly efficient optimization algorithms without spending time on it. On the other hand any kind of model modification, that requires direct access to the solver, becomes unfeasible. This is especially relevant for attempts to increase model performance or to analyze model outputs (e.g. uncertainty analysis). Another aspect of GAMS is, that its syntax is kept quite simple, which allows it to be learned fast. This even holds true for people, that are not used to work with programming languages. Unfortunately, this feature is bought by a quite limited versatility of the language. Many problems, such as dynamic sizes of vectors, or advanced mathematical calculations, can only be done in an circuitous and unsatisfactory way or cannot be performed with GAMS at all. This led in the case of MAgPIE to an outsourcing of many model components to a collection of scripts written in R [R Development Core Team, 2010], PHP [Bakken et al., 2000] and Python [van Rossum and Drake Jr., 2001] embracing the GAMS core. One example of an outsourced component is the clustering and further preprocessing of high-resolution model inputs described in chapter 3. As solver MAgPIE currently uses CONOPT [Drud, 1994], a solver especially designed for large, nonlinear problems.

2.6 Outline of this thesis

To allow policy makers to take reasonable decisions it is an urgent need, that scientists deliver information about processes involved in the respective topic. One approach to estimate and predict the role and impacts of processes is the development of models represesenting the sector of interest. The basic duty of models is to identify the interactions, that occur between different processes. A typical setup is: We know process A and we know process B, but what do we get, if we combine both processes? Applied to agriculture this can mean: We know the demand for several agricultural commodities at the market level and we know the spatial yield variations of these agricultural commodities. What production patterns do we get, if we combine both findings? In most cases this combination of processes comes along with the problem of different scales. In the

mentioned example the agricultural yield is related to a local scale. It depends on local factors such as water availability, weather and soil conditions. In contrast the demand for an agricultural commodity is described at the market level at which the products are traded. Combining these processes, all involved scales must be taken into account by the model. Having only a local model the yield could be computed correctly, but the demand could not be estimated properly. Having only a model acting at the market scale, the demand estimation would be more reasonable, but relevant yield variations could not be taken into account. Combining all relevant scales becomes a challenging task, especially when many different scales are involved. Global land-use models, such as MAgPIE, are prime examples of models facing these scale issues. While some processes occur at the global scale, such as global trade, others are of a regional character, such as technological development. The lower end is marked by processes occurring at the local scale, such as farming decisions and plant growth. The ability to properly include all these scales into a model is strongly limited by the availability of computational resources. The only chance to get an accurate model, which is still computable, is therefore to condense these cross-scale interactions effectively. Accordingly, the overall research question of my thesis is: How can a condensation of cross-scale interactions in a land-use model be achieved?

Applied on an existing model one can split this question into two parts: First, how can existing cross-scale interactions be simplified and optimized? Second, how can important cross-scale linkages be established, which are missing in the existing framework? The answers to both questions significantly depend on the involved processes and boundary conditions, such as the general model setup or the kind of information, which is available for the corresponding topic. This eliminates the potential to give a generalized answer. Therefore, I apply these research questions on the MAgPIE model and present solutions for specific cross-scale interactions, which are of high importance for the general model simulation. In this context both cases, the improvement of existing cross-scale linkages and the establishment of missing linkages, are treated exemplary.

Chapter 3 deals with the issue how to optimize existing cross-scale implementations. It is focused on the issue of data upscaling and its central question is: How can the amount of information, which is lost in the upscaling process, be reduced? A selection of clustering methods (k-means and hierarchical clustering) is applied for high-resolution MAgPIE input data and compared to the common upscaling method using a static grid, which is a upscaling rule independent of the underlying data. To estimate the quality of the different methods and the corresponding resolution losses, two comparisons are performed: First, the similarity between the original input data and its corresponding upscaled data is measured and compared between the different upscaling methods. This is used as an indicator for the general information loss in the upscaling procedure. Second, the upscaled data is used as input for simulation runs with MAgPIE and the corresponding model outputs are compared with model outputs derived with the original, high-resolution input data. This comparison provides insights concerning the loss of information which is most relevant for the model simulation itself.

Chapter 4 and 5 deal with the aspect of improving but also establishing important cross-scale linkages in a model. The central question for these two chapter is: How can

technological change be implemented properly in a global land-use model? Technological change acts on two scales: Investments are done and research takes place at the regional scale, whereas its outcome can be observed at the local scale in form of yield increases and proposed management changes. As I started to work with MAgPIE this linking was already - in contrast to most other agricultural land-use models - present in the model, so that it was possible to induce yield increases due to investments in technological change. However, technological change is a bidirectional cross-scale interaction: Regional investments in technological change can trigger yield increases at the local scale, but the technological development level at the local scale also affects investment effectiveness (investment-yield ratio) at the regional scale. The required investigations for this missing feedback link are split in two parts. In the first part a measure for agricultural land-use intensity is developed. It measures the current level of yield increases due to past technological developments in a simple and condensed manner (Chapter 4). This measure is necessary for the mathematical representation of agricultural development in each region at each timestep. In the second part this measure is used to investigate the dependence of the investment-yield ratio (which is describing how much investment is required for a certain amount of yield increase) on the current level of agricultural land-use intensities. Thereafter, this relation is utilized for a condensed, bidirectional implementation of technological change in MAgPIE (Chapter 5).

3 Cluster-based upscaling

Abstract

[1] Global land-use models have to deal with several spatial scales, ranging from the global scale of $10^5 km$ down to the farm level with scales of around $100m$. Combined with the increasing complexity of modern land-use models one faces the problem of limited computational resources. One solution of this problem is a spatial upscaling based on a static grid or administrative units as e.g. countries. Unfortunately this type of upscaling flattens many regional differences and produces a homogenized map of the world. In this chapter I present an alternative upscaling approach using clustering methods. Clustering reduces the loss of information due to upscaling by choosing an appropriate aggregation pattern.

In the following different clustering methods are investigated concerning their quality in terms of information conservation. The results indicate that clustering is always a good choice and preferable compared to grid-based upscaling. Although all tested clustering methods delivered better results than grid-based upscaling, the choice of clustering method is not arbitrary. Comparing original and upscaled data directly shows, that k-means clustering is the best choice using a small number of clusters, while bottom-up hierarchical clustering behaves best for light upscaling with a number of clusters close to the original number of cells. Comparing outputs of a model fed with original data and a model fed with upscaled data, bottom-up clustering delivered the best results for all numbers of clusters tested (ranging from 14 to 3772 clustered of a data set containing 4669 cells).

3.1 Introduction

One characteristic of land-use models is their linking of elements from geography and economics. Since the general approaches of both disciplines differ significantly several scale related problems arise: In geography spatial information plays a major role. Data is linked to a location and spatial explicitness is most desirable. In economics markets and market equilibria are key elements. Spatial explicitness typically plays only a minor role. Instead the focus lies on complex market dynamics and flows of inputs and outputs.

The challenge of agricultural land-use models is to take both aspects into account: global markets and their market equilibria as one important feature of the agricultural sector, but also spatial explicitness of agricultural production, since productivity in agriculture strongly depends on local environmental conditions. However, including high-resolution data into an equilibrium model leads to significant problems of computability. Increasing the number of simulated units typically leads to disproportionate increases

[1] This chapter is based on Dietrich et al. [2011a]

in computation time and required amount of working memory. For instance, MAgPIE (Chapter 2.3) shows quadratic increases in computation time with increasing number of simulated cells. So a bisection of the grain side length, which means a quadrupling of 2D-cells, leads to 16 times longer computation times. Furthermore, the increases in working memory requirements limit the total number of cells to less than 5000.

In current agricultural research different approaches are used to deal with this problem. Models focused on the economy often cover global agricultural markets, but only at a coarse spatial resolution of a few world regions (e.g. AgLU [Sands and Leimbach, 2003], FASOM [Adams et al., 1996], IMPACT [Rosegrant et al., 2008]), whereas models focused on geographical or ecological processes are either only modeling certain regions of the world, with exogenous global markets (e.g. CLUE [Verburg et al., 1999a,b, Wassenaar et al., 2007]), or apply a rule-based approach (e.g. SALU [Stephenne and Lambin, 2001], Syndromes [Cassel-Gintz and Petschel-Held, 2000] - a general land use model review was done by Heistermann et al. [2006]). Hence, either the economic or the ecological part is represented in a simplified manner bypassing or disregarding the issue of different scales.

One possibility to cope with this issue of cross-scale interactions is the use of cluster algorithms for upscaling. In the following I present and compare a selection of clustering algorithms as methods to increase the spatial resolution of agricultural equilibrium models. For the comparison the MAgPIE model is used. First, I have modified the model structure to be able to simulate in any spatial aggregation, with a lower limit at the grain size of the original input data. Second, I have implemented spatial upscaling methods (grid-based and clustering-based) to merge input data to these aggregations (together this allows running the model at various spatial aggregations). Third, I have implemented an interpolation methodology to downscale clustered outputs back to the grain size of the input data. Last, I have used this implementation to compare the standard upscaling method using a static grid with hierarchical and non-hierarchical clustering methods.

3.2 Methods

3.2.1 Model implementation

As described in chapter 2.3, MAgPIE is a model with three scales involved: A global scale representing global markets, a regional scale of 10 world regions representing specific economic development, demands and technology levels, and a local scale representing farming decisions based on spatially varying production parameters, as for instance potential yields and water availability. Since GAMS does not allow for calculating sets and therefore cannot handle inputs in varying resolutions, a PHP script is executed before GAMS is started (Chapter 2.5). The PHP script organizes the upscaling of the original input data set and rewrites the sets in the GAMS source code according to the chosen aggregation. The upscaling of input data itself is done in R, either by using a static or a clustering grid. After finished execution of the GAMS model, the clustered data is downscaled to the grain size and grid of the original input data using another R and Python script.

The unprocessed input data has a grain size of 0.5° (i.e. 30 arc-minutes longi- and latitudinal). Each cell contains information on potential yields of 20 different crops (rainfed and irrigated)[2], crop-specific demands for irrigation water, the total amount of water available for irrigation (all calculated with LPJmL - Chapter 2.4), total cropland area and total land available for additional cropland expansion [Krause et al., 2009]. For upscaling two approaches are implemented: (A) an upscaling based on static grids and (B) an upscaling using clustering methods. In any case only cells that belong to the same world region are aggregated together.

In the case of static grids a grain size in degree is chosen (coarser than the original grain size of 0.5°) and input data cells belonging to the same cluster are either summed up or (weighted) averaged dependent on the type of data. Yields are averaged using the total crop share of a cell as weight; the amount of available water per cell is summed up; required amount of water for each crop is also crop-area weighted averaged; and crop shares are cell-area weighted averaged.

For the clustering methods the target grid is chosen depending on the data to be upscaled. All clustering methods have in common that clusters are built on some kind of distance information between cells/clusters. Every cell is represented by its data and the distance between cells is based on the similarity of data, for instance cells with similar yields are close to each other, whereas big differences in yields lead to high distances between cells (not to be mistaken for physical distance). Because of regional separation every cluster belongs exactly to one region. In contrast to grid-based upscaling, clusters are not connected to a well defined spatial location. It can even happen that one cluster is split in several fractions distributed over the whole region. Furthermore, clustering does not increase the grain size, since the smallest unit, which a clustered data set can contain, is still one cell of the original data set. Instead of increasing the grain size, cluster methods try to reduce the amount of resulting units by combining cells with similar characteristics.

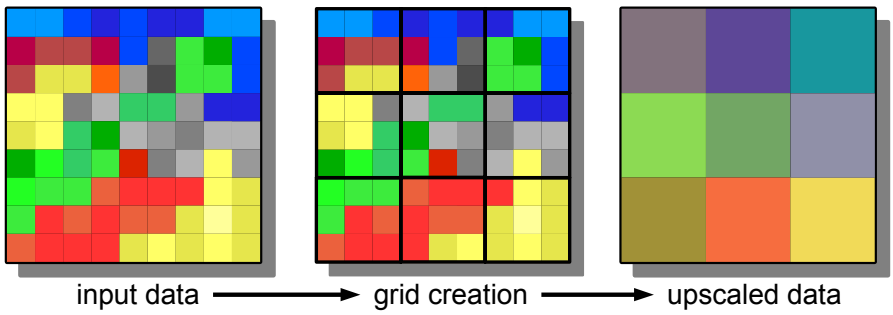

Figure 3.1: Schematic diagram showing the upscaling process using a static grid. 81 cells are upscaled to 9 data elements.

Figure 3.1 and Figure 3.2 illustrate the different upscaling approaches schematically.

[2] wheat, rice, maize, millet, pulses, cotton, potato, sugar beet, sugar cane, cassava, sunflower, soybean, groundnut, palm oil, rapeseed, bioenergy grasses, bioenergy trees, fodder, pasture, others

Using a static grid the procedure is quite simple (Figure 3.1). In the shown example the initial data set has 81 cells with a grain size of 1x1 and an extend of 9x9. The values of each cell are indicated with colors. Increasing the grain size to 3x3 leads to the static grid, which is used for upscaling. The colors of the new data set are derived by averaging the colors of all elements within a segment. Depending on the homogeneity of a segment the resulting color is either similar to the colors of the input data set (lower-right segment) or quite different (upper-left segment). In the shown example the initial color distribution is only hardly visible in the upscaled diagram, which indicates a significant loss of resolution.

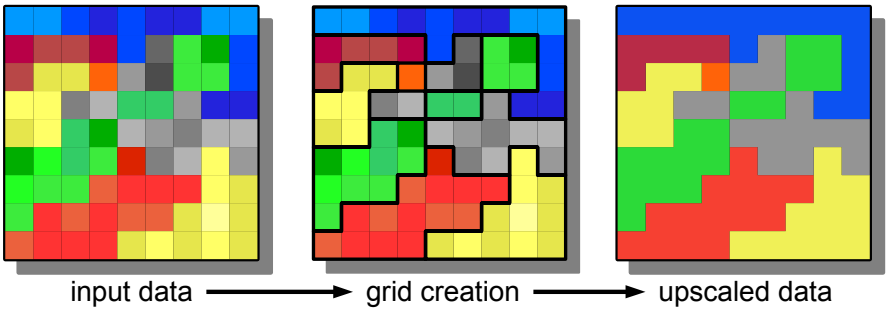

Figure 3.2: Schematic diagram showing the upscaling process using clustering techniques. 81 cells are upscaled to 7 cluster.

Figure 3.2 schematically shows cluster-based upscaling starting with the same input data set as in figure 3.1. Using clustering the grid is determined based on the input data itself. In the grid creation step, cells are merged based on its similarity. In contrast to static grid upscaling the cluster size is variable. This allows to build clusters with only one cell (orange cell in the schematic diagram). The grain size of the data set is not increased. Merged cells do not have to be neighbors. A distribution of one cluster over the whole data set is possible (e.g. the green or the yellow cluster). However, this eliminates the possibility to reasonably link a cluster with a single, spatial coordinate. Therefore, the upscaled data typically has to be downscaled again, before it can be used for spatially explicit calculations. The deviation between the input and the upscaled data set shows, that some data is lost. However, the characteristics of the initial data set are still visible, indicating a relatively high level of resolution conservation. In this context it is also interesting to note, that the shown cluster example only used 7 cluster, while the static grid upscaling was performed with 9 segmentation units.

3.2.2 Clustering

Two clustering methods are used in several variations: k-means clustering and hierarchical clustering. Using many dimensions (meaning many data sets involved) for clustering will slow down the process and decrease the similarity of clustered cells concerning a specific property, as e.g. the rainfed maize yield. On the other hand excluding data from the clustering process can produce significant biases. To decide which values should be in-

cluded in the clustering I distinguished between two types of values: Values with spatial inhomogeneity that significantly influence decisions made by the model and values with spatial inhomogeneity that has no significant impact on the simulation. One example for a significant value is the maize yield in each cell. Increasing the inhomogeneity in yields also has an effect on farming decisions: An increase in inhomogeneity increases the degree of specialization in an optimization model since one cell becomes more favorable than another. An increased homogeneity of yields reverses this effect, suitability for production becomes more similar between clusters. As the degree of specialization is also important at the global scale due to the described nonlinear cross-scale interaction, this kind of information is relevant for the simulation. Typically a higher degree of specialization decreases overall production costs. So averaging highly inhomogeneous yields should increase total production costs of the model, since potential for optimization is lost.

An example for a non-significant variable is total cropland area. An increased inhomogeneity does not change general farming decisions. Land availability does not increase or decrease only because of a homogeneous or inhomogeneous distribution. The degree of specialization will remain the same, independent of the cropland homogeneity. At least in first order the general model behavior is not influenced by the homogeneity of cropland. For cropland area it is relatively obvious that its spatial inhomogeneity does not affect the simulation significantly, but in other cases this distinction becomes much more problematic. One may look, for example, at the irrigated yield of cotton. This value can become important in some clusters, but in most cases it should be irrelevant because the production of irrigated cotton will stay unattractive for a broad range of yields. In this case one can expect to produce errors by excluding it from the clustering process but it is not clear which type of error will be larger: the error caused by the exclusion or the error that is caused by the fact, that the cotton yield will change the clustering in many regions where cotton itself is irrelevant and therefore, should not influence the clustering. Since, except of total cropland shares, no data set could be clearly identified as non-significant all remaining data was used for the comparison of the different upscaling types.

k-means clustering

k-means is a method to partition n cells with data $x_1, ..., x_n$ into k clusters $\mathbf{S}_1, ..., \mathbf{S}_k$ with mean values $\mu_1, ..., \mu_k$ by minimizing the within-cluster sum of squares (WCSS) (Equation 3.1) [Hastie et al., 2005, Hartigan, 1975].

$$\arg_{\mathbf{S}} \min \sum_{i=1}^{k} \sum_{x_j \in \mathbf{S}_i} \|x_j - \mu_i\|^2 \tag{3.1}$$

The problem can be solved analytically, but because of its high computational intensity typically heuristic algorithms are used. I have used the heuristic implementation offered by the pycluster python module [de Hoon et al., 2004] which is based on the EM-algorithm. It is an iterative algorithm which can be divided into three steps:

1. Calculate the centroid (average value) of each cluster
2. For each item, determine which centroid is the closest
3. Reassign the item to that cluster

For initialization cluster centroids are distributed stochastically. The iteration is stopped, when no items are reassigned anymore. This approach is relatively fast and delivers accurate results. A disadvantage is that the algorithm is not fully deterministic, because of its initialization. Therefore, reproducibility is limited, which can be problematic when comparing results of different runs. Furthermore, comparison of runs with different numbers of clusters is impracticable because of its missing cluster hierarchy.

In the case of MAgPIE one has to face the additional constraint that all cells within one cluster have to belong to the same world region. For k-means clustering this is assured by initializing separate clustering processes for each region. The number of clusters for each region is determined by multiplying the fraction of cells within a region relative to total cells with the number of total clusters.

Hierarchical clustering

Hierarchical clustering methods can be classified in two types: top-down and bottom-up. Both create a hierarchy based on the distances between cells/clusters. Top-down methods start with one cluster which is split step by step into smaller clusters until every cluster contains exactly one cell. Bottom-up methods start on the cell level and aggregate cells step by step to clusters until only one cluster remains. The distances between clusters can be measured with various metrics and methods. I have used an euclidean metric and the maximal-linkage method, which is measuring the distance between clusters as the maximum distance between two cells of both clusters. Single linkage (minimum distance between cells of two clusters), average linkage (average distance between cells of both clusters) and centroid linkage (distance between centroids of two clusters) were also tested for bottom-up clustering. However, I excluded them, since they produced at best equivalent but in most cases significantly inferior results compared to maximal-linkage in terms of the used evaluation measures (Section 3.2.4).

Hierarchical clustering has the advantages that it is fully deterministic and that its results produced with different numbers of clusters can be merged easily based on its hierarchy. However, the hierarchy also limits the adaptivity of the clusters to the data, because of lock-in effects: Clustering decisions made earlier in the hierarchy cannot be reversed. Hence, my assumption is that top-down clustering becomes ineffective for many clusters whereas bottom-up clustering shows the same effect for few clusters. For comparison I use a bottom-up hierarchical clustering as implemented in the pycluster module [de Hoon et al., 2004] and an own implementation of top-down clustering also based on the maximal-linkage method.

The combination of two cells or clusters to a new cluster in the bottom-up approach is well-defined, whereas the partition of one cluster into sub-clusters or cells in the

top-down approach can be done in many ways. This condition makes top-down clustering less trivial and less adaptive compared to bottom-up. For instance, single-linkage and average-linkage approaches applied for top-down clustering become computationally highly intensive and hard to apply for large datasets. Only the maximal-linkage approach can be applied in a straightforward way and is therefore implemented for the comparison. First of all, a distance matrix is calculated containing all distances between cells within the cluster, then this matrix is sorted by distances and iterated starting with the longest distance. The first two cells in the sorted matrix are separated directly and put to clusters 1 and 2. Starting with the second cell-linkage of the distance matrix, it is tried to separate as many cells with long distances as possible: When a cell is already mapped to a cluster the other cell is put into the opposite cluster. If no cell is mapped to a cluster, both cells are mapped to two temporary clusters. If one cell belongs to an output cluster and one to a temporary cluster, both connected temporary clusters are merged with the output clusters. No action is taken if both cells are already mapped to an output cluster. This procedure assures that as many long-distance connections as possible are cut and that both cells with the longest distances between them belong two different clusters (Appendix 1).

Hierarchical clustering results in a hierarchy tree which is describing in which order cells are clustered based on the distance. To prevent clustering across region borders I started separate clustering processes for each region. However, in contrast to k-means clustering, I merged the results of all ten world regions to one hierarchy tree based on distance information (clusters with smallest distances are combined first). This approach allows for distributing clusters in a globally optimal way, while preventing cross-regional clustering.

3.2.3 Downscaling

Two significant problems arise, when dealing with clustered data. First, the clustered data does not necessarily have a spatial meaning, since clusters can be distributed over the whole map. Second, only outputs directly derived from input data involved in the clustering procedure have a proper meaning. With respect to values that were not part of the clustering, model outputs in each cluster are only the mean of a potentially strongly inhomogeneously distributed value and a significant amount of information is lost. Both problems can be solved with downscaling. The problem of missing spatial information is solved by inversion of the upscaling process, whereas the information loss in the second case is compensated by additional knowledge about the distribution used within the downscaling process.

The downscaling can be divided into two steps. In step one, data with trivial downscaling rules is processed. Trivial downscaling is either downscaling by giving each cell the value of its belonging cluster or downscaling by partitioning a summed-up value proportionate to data that is known already in the higher resolution. In all other cases downscaling rules are non-trivial. Values, which require non-trivial downscaling, are not processed directly. Instead, these values are split into components with trivial downscaling rules and recalculated in a second step at 0.5° grain size level based on these

downscaled components.

One example is cellular, crop-specific data on cropland shares. First, these shares are split into a term containing crop-specific total areas and a term containing the total cluster areas. Second, crop-specific total areas are downscaled by partitioning it proportionally to total cell area (total areas are constant and known at 0.5° grain size and, therefore, do not need to be downscaled). Third, the crop-specific areas are divided by the total area to reconstruct the original shares at 0.5° grain size.

3.2.4 Evaluation

The suitability of a clustering method strongly depends on the given task. There is no "best" clustering algorithm in general [Hartigan, 1985, Jain and Dubes, 1988, pg. 142]. While some methods appear to be more often an appropriate choice than others, there is no general ranking of the different approaches. To test the quality of the different upscaling methods applied to the task given in this dissertation I have upscaled the original 0.5° data with 59199 cells to a grain size of 2.0° with 4669 cells. This new data was then upscaled to 17 different aggregation levels with 3772 to 14 clusters using static grids, k-means clustering, hierarchical bottom-up clustering and hierarchical top-down clustering. Afterwards the upscaled data and the related model output was downscaled again by giving each 2.0° cell the values of its related cluster. In a second experiment the same procedure was repeated holding the upscaling method fixed to hierarchical bottom-up clustering with 400 and 1438 cells but changing this time the number and type of data sets involved in the clustering. Besides the standard case of using all available data except of cropland shares, also upscaling with each data set as single input and upscaling with the 10 most relevant crop and 10 least relevant crops (based on quality measures of single input upscalings) was performed.

To measure the quality of the different upscalings $i = 1..n$, two measures $d_1(i)$ and $d_2(i)$ are applied. Both measures can be written in general as the mean of normalized distances (Equation 3.2).

$$d_k(i) = \frac{1}{m} \sum_{s=1}^{m} \frac{\hat{d}_k(X_{i,s}, X_{0,s})}{\max_{j=1}^{n} \hat{d}_k(X_{j,s}, X_{0,s})} \tag{3.2}$$

The quality of the i-th upscaling is calculated by taking the mean of the distances $\hat{d}_k$ between all data sets $s = 1..m$ used for comparison (e.g. rainfed yield of maize is one data set, available discharge of each cell is another one), normalized by the maximum distance observed for each data set. This approach delivers values between 0 and 1, with 0 in the case of a perfect match between reference and upscaled data and 1 in the case of the maximum observed deviations between data sets. The measures can be interpreted as measures for the resolution of a data set, where values close to 0 indicate a resolution similar to the original data set and values close to 1 indicate high losses in resolution. For the calculation two distance measures $\hat{d}_1$ and $\hat{d}_2$ were applied:

1. The euclidean distance $\hat{d}_1$ between two data sets Y and Z with l data points:

$$\hat{d}_1(Y,Z) = \|Y - Z\| = \sqrt{\sum_{k=1}^{l} (y_k - z_k)^2} \tag{3.3}$$

   This measures the similarity between original and upscaled data set: the more information is conserved, the lower the final value will be.

2. The mutual information distance d_2. For calculation of mutual information distances I have used the R package "bioDist"[Gentleman et al., 2005]. Mutual information m is a nonlinear measure for the mutual dependence of two variables Y and Z (Equation 3.4).

$$m(Y,Z) = \sum_{y \in Y} \sum_{z \in Z} p(y,z) \log \left(\frac{p(y,z)}{p(y)p(z)} \right) \tag{3.4}$$

   $p(y,z)$ is the joint probability density function of Y and Z, $p(y)$ and $p(z)$ are the marginal probability density functions of Y and Z respectively. Mutual information is a measure for the amount of information which is shared by both variables. So it should be high if the upscaling conserves much information of the original dataset and should be low otherwise. The mutual information distance $\hat{d}_2$ is calculated by applying the transformation of equation 3.5 to the mutual information m, which was proposed by Joe [1989].

$$\hat{d}_2(Y,Z) = 1 - \sqrt{1 - \exp(-2m(Y,Z))} \tag{3.5}$$

   Like $\hat{d}_1$, $\hat{d}_2$ is also measuring the similarity between original and upscaled data set.

description	**unit**
cellular, crop-specific, annual water demand required for optimal irrigation	$mm/year$
cellular, annual amount of water available for irrigation	$10^6 m^3$
cellular, crop-specific rainfed yields	ton/ha
cellular, crop-specific yields under optimal irrigation	ton/ha

Table 3.1: List of cellular model inputs

Measures are applied separately on cellular model inputs, which are also involved in the clustering procedure (Table 3.1), and a selection of default non-cellular and cellular model outputs (Table 3.2, Table 3.3). Since the mutual information measure d_2 does

description	unit
total costs of production	$10^6 US\$$
total production value (price·amount)	$10^6 US\$$
global area of all crops	$10^6 ha$
gross global area of converted land	$10^6 ha$
regional area of all crops	$10^6 ha$
global area-weighted average of technical change rates	(dimensionless)
global supply-demand balances	(dimensionless)
technical change rates in regions	(dimensionless)

Table 3.2: List of non-cellular model outputs

description	unit
cellular land use shares of single crops in total area	(dimensionless)
cellular land use shares of total cropland, pasture and land available for cropland expansion	(dimensionless)
cellular cropland prices	$US\$/ha$
cellular producer rent including land rent	$10^6 US\$$

Table 3.3: List of cellular model outputs

require time series, it is only applied on input data sets and cellular output data sets, whereas d_1 is also applied on non-cellular data sets.

To test the significance of measured differences in quality between upscaling types the Wilcoxon signed-rank test is used [Wilcoxon, 1945].

3.3 Results

3.3.1 Comparison of cluster methods and number of clusters

Applying the measures d_1 and d_2 and a Wilcoxon signed-rank test with H_0 hypothesis dataset1 $\geq$ dataset2 the clustering methods show for the model input data significantly better results over the full tested range compared to the standard upscaling using static grids (Figure 3.3, Table 3.4). Overall k-means shows the best results in the case of strong upscaling for which it behaves significantly better than both hierarchical methods. In the middle range k-means still delivers significantly better results compared to top-down hierarchical clusters, but non-significant advantages compared to bottom-up clustering. For light upscaling bottom-up clustering shows the best performance, which is significantly better than top-down clustering in both quality measures and significantly better than k-means clustering in d_1 (Table 3.4). Top-down clustering behaves over the whole

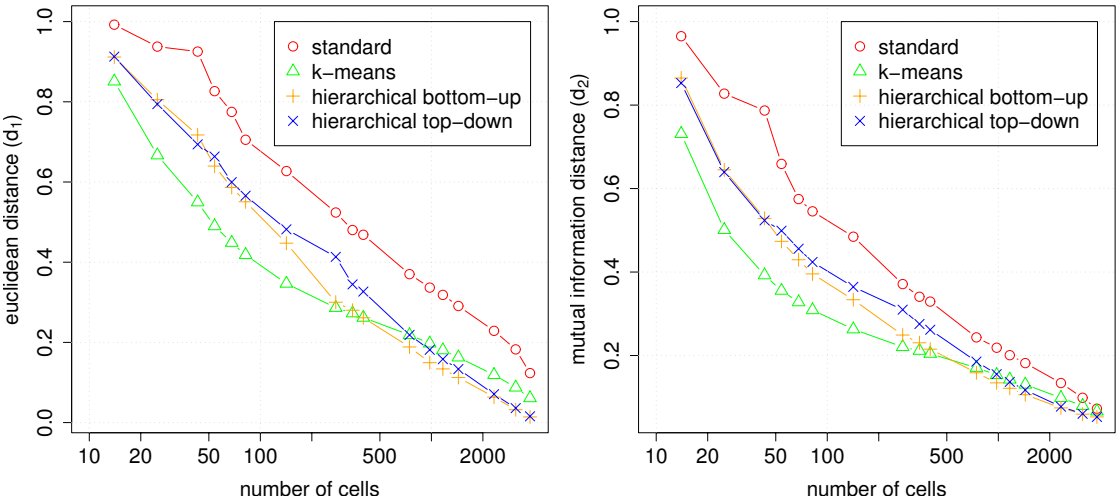

Figure 3.3: Upscaling quality measured with d_1 (left) and d_2 (right) on upscaled model input data.

range slightly worse than bottom-up clustering, with significant differences for medium and light upscaling and nearly identical results for strong upscaling.

Comparing the outputs of model runs with reference data and with upscaled data the picture changes (Figure 3.4, Table 3.5). Still all three clustering methods perform significantly better than upscaling using static grids over the whole range, but with lower significance levels and smaller differences in quality measures. Whereas the quality ratings of input data varied nearly over the whole range from 0 to 1, the ratings of output data only range from 0.3 for d_1 or 0.5 for d_2 to 0.9. This reduced variance in quality measures reflects an increased variance in rankings of the different upscaling experiments across different output data sets. That means the distance rankings of different upscaling methods and number of clusters depend more on the chosen set of output data, whereas the rankings in input data primarily depend on the number of clusters and upscaling method. Hence, some outputs might get better with k-means clustering, other outputs with hierarchical clustering. It even happens that a less aggressive upscaling increases the distance between the reference and the upscaled case.

Comparing the output results of the different clustering methods with input results one finds, that k-means performs much worse in case of the outputs. On the input side there is a significant quality increase especially for strong upscaling, when applying k-means clustering compared to hierarchical clustering. Whereas d_2 still reports significantly better results for k-means at strong upscalings, this effect cannot be observed in outputs for d_1 anymore. In fact k-means shows in d_1 for strong upscalings an even worse behavior then static grids. Comparing the hierarchical clustering methods the results of the input data persists also for the output case: Both methods produce similar results with slight, but significant advantages for the bottom-up approach (significant for medium to low upscaling levels, non-significant for strong upscaling).

Analyzing non-cellular and cellular outputs separately one finds, that in general results

H_0		full		coarse		medium		fine	
h ≥ s	d1	0.000	***	0.004	**	0.001	***	0.001	***
	d2	0.000	***	0.004	**	0.001	***	0.001	***
k ≥ s	d1	0.000	***	0.004	**	0.001	***	0.001	***
	d2	0.000	***	0.004	**	0.001	***	0.001	***
t ≥ s	d1	0.000	***	0.004	**	0.001	***	0.001	***
	d2	0.000	***	0.004	**	0.001	***	0.001	***
s ≥ h	d1	1.000		1.000		1.000		1.000	
	d2	1.000		1.000		1.000		1.000	
k ≥ h	d1	0.112		0.004	**	0.138		0.993	
	d2	0.044	*	0.004	**	0.053		0.903	
t ≥ h	d1	0.999		0.945		1.000		1.000	
	d2	0.999		0.961		1.000		0.999	
s ≥ k	d1	1.000		1.000		1.000		1.000	
	d2	1.000		1.000		1.000		1.000	
h ≥ k	d1	0.897		1.000		0.884		0.010	**
	d2	0.960		1.000		0.958		0.116	
t ≥ k	d1	0.993		1.000		0.993		0.539	
	d2	0.997		1.000		0.998		0.722	
s ≥ t	d1	1.000		1.000		1.000		1.000	
	d2	1.000		1.000		1.000		1.000	
h ≥ t	d1	0.001	***	0.074		0.001	***	0.001	***
	d2	0.002	**	0.055		0.001	***	0.002	**
k ≥ t	d1	0.009	**	0.004	**	0.010	**	0.500	
	d2	0.004	**	0.004	**	0.003	**	0.312	

Table 3.4: Input data quality comparison: p-values and related significance levels of a Wilcoxon signed-rank test applied to the d_1 and d_2 results of upscaled model inputs (* p≥95%, ** p≥99%, *** p≥99.9% | h: hierarchical bottom-up, t: hierarchical top-down, k: k-means, s: static grid | full: full range (14-3772 cells), coarse: strong upscaling (14-346 cells), medium: medium upscaling (54-1167 cells), fine: light upscaling (346-3772 cells))

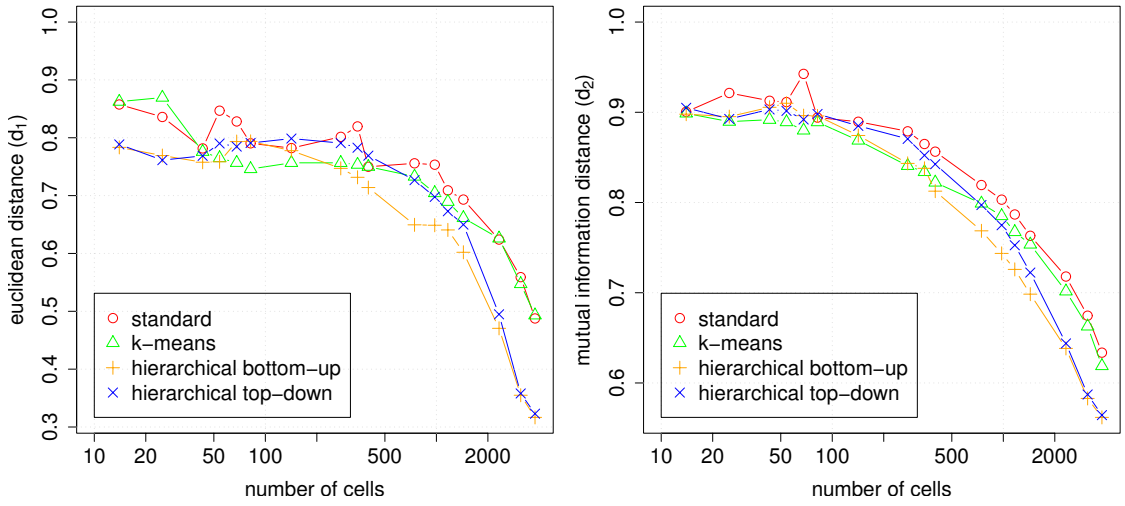

Figure 3.4: Upscaling quality measured with d_1 (left) and d_2 (right) on model outputs derived with upscaled model inputs.

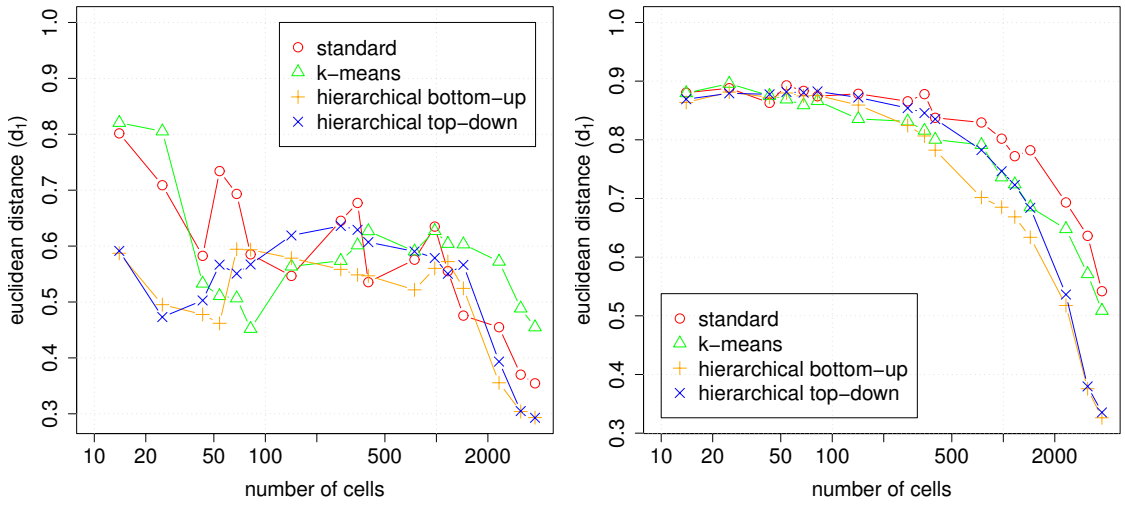

Figure 3.5: Comparison of upscaled results using quality measure d_1 for non-cellular model outputs (left) and cellular model outputs (right).

H_0		full		coarse		medium		fine	
h ≥ s	d1	0.000	***	0.008	**	0.002	**	0.001	***
	d2	0.000	***	0.012	*	0.003	**	0.001	***
k ≥ s	d1	0.003	**	0.039	*	0.002	**	0.014	*
	d2	0.000	***	0.004	**	0.001	***	0.001	***
t ≥ s	d1	0.000	***	0.039	*	0.024	*	0.003	**
	d2	0.000	***	0.020	*	0.002	**	0.001	***
s ≥ h	d1	1.000		0.996		0.999		1.000	
	d2	1.000		0.992		0.998		1.000	
k ≥ h	d1	0.998		0.680		0.884		1.000	
	d2	0.920		0.008	**	0.688		0.997	
t ≥ h	d1	1.000		0.926		0.997		1.000	
	d2	0.998		0.680		0.993		1.000	
s ≥ k	d1	0.997		0.973		0.999		0.990	
	d2	1.000		1.000		1.000		1.000	
h ≥ k	d1	0.002	**	0.371		0.138		0.001	***
	d2	0.087		0.996		0.348		0.005	**
t ≥ k	d1	0.274		0.629		0.990		0.188	
	d2	0.500		1.000		0.968		0.116	
s ≥ t	d1	1.000		0.973		0.981		0.998	
	d2	1.000		0.988		0.999		1.000	
h ≥ t	d1	0.001	***	0.098		0.005	**	0.001	***
	d2	0.003	**	0.371		0.010	**	0.001	***
k ≥ t	d1	0.741		0.422		0.014	*	0.839	
	d2	0.518		0.004	**	0.042	*	0.903	

Table 3.5: Output data quality comparison: p-values and related significance levels of a Wilcoxon signed-rank applied to the d_1 and d_2 results of model outputs generated with upscaled inputs (* p≥95%, ** p≥99%, *** p≥99.9% | h: hierarchical bottom-up, t: hierarchical top-down, k: k-means, s: static grid | full: full range (14-3772 cells), coarse: strong upscaling (14-346 cells), medium: medium upscaling (54-1167 cells), fine: light upscaling (346-3772 cells))

H_0		full		coarse		medium		fine	
h ≥ s	non-cellular	0.001	**	0.020	*	0.042	*	0.014	*
	cellular	0.000	***	0.039	*	0.002	**	0.001	***
k ≥ s	non-cellular	0.627		0.098		0.138		0.968	
	cellular	0.000	***	0.055		0.001	***	0.001	***
t ≥ s	non-cellular	0.032	*	0.020	*	0.188		0.312	
	cellular	0.001	***	0.191		0.007	**	0.001	***
s ≥ h	non-cellular	0.999		0.988		0.968		0.990	
	cellular	1.000		0.973		0.999		1.000	
k ≥ h	non-cellular	0.995		0.809		0.784		1.000	
	cellular	0.994		0.371		0.812		1.000	
t ≥ h	non-cellular	0.985		0.809		0.968		0.993	
	cellular	1.000		0.992		1.000		1.000	
s ≥ k	non-cellular	0.391		0.926		0.884		0.042	*
	cellular	1.000		0.961		1.000		1.000	
h ≥ k	non-cellular	0.005	**	0.230		0.246		0.001	***
	cellular	0.007	**	0.680		0.216		0.001	***
t ≥ k	non-cellular	0.122		0.629		0.920		0.042	*
	cellular	0.627		0.945		0.997		0.312	
s ≥ t	non-cellular	0.972		0.988		0.839		0.722	
	cellular	0.999		0.844		0.995		1.000	
h ≥ t	non-cellular	0.017	*	0.230		0.042	*	0.010	**
	cellular	0.000	***	0.012	*	0.001	***	0.001	***
k ≥ t	non-cellular	0.888		0.422		0.097		0.968	
	cellular	0.391		0.074		0.005	**	0.722	

Table 3.6: Comparison between cellular and non-cellular outputs: p-values and related significance levels of a Wilcoxon signed-rank applied to the d_1 results of cellular and non-cellular model outputs (* p≥95%, ** p≥99%, *** p≥99.9% | h: hierarchical bottom-up, t: hierarchical top-down, k: k-means, s: static grid | full: full range (14-3772 cells), coarse: strong upscaling (14-346 cells), medium: medium upscaling (54-1167 cells), fine: light upscaling (346-3772 cells))

for non-cellular data are much fuzzier than for cellular ones (Figure 3.5, Table 3.6). For hierarchical clustering methods both outputs still show significant quality improvements compared to static grids. However, for k-means the results only remain significant for cellular data. For non-cellular data static grids deliver even better results than k-means in the case of light upscaling.

3.3.2 Choice of data sets involved in clustering

		input				**output**			
		d_1		d_2		d_1		d_2	
		h400	h1438	h400	h1438	h400	h1438	h400	h1438
all data		1	1	1	1	1	2	1	2
input	**best**	3	3	3	3	35	33	5	4
	worst	35	14	5	5	5	8	3	3
output	**best**	4	4	4	4	6	1	2	1
	worst	2	2	2	2	9	24	4	5

Table 3.7: Ranks concerning input and output quality of selected upscalings using hierarchical bottom-up clustering with 400 (h400) and 1438 (h1438) clusters and varying numbers and types of data sets used for upscaling: "all data" = All available data sets, "input/output best/worst": A combination of those 10 data sets showing the best/worst performance in input/output quality measures, when only using that single data set for upscaling. The ranks are based on an quality list containing the shown 5 combinations plus all 60 single data set upscalings (Appendix - Table 1).

Table 3.7 shows the rank of five different data set combinations used for upscaling concerning their input and output quality measures out of a list of 65 data set combinations in total (the 5 shown upscalings + 60 single data set upscalings). This test was performed in order to estimate the role of data set choice for upscaling.

Results show, that especially the number of data sets involved in the upscaling procedure matters. More data sets deliver better results in general. Hence, the "all data" upscaling, which was also applied for the previous tests, delivers the best results in total. But also the tests using only the 10 most important data sets (best) and 10 least important (worst) delivered in most cases better results than any other run using only a single data set for upscaling.

Comparing the best and worst cases one finds, that in the case where not all data sets are used, still the specific choice of data sets matters. Some data sets deliver better patterns than others. But having a data set that delivers good quality results for the upscaled input data does not necessarily mean the model outputs will also have a good quality. In fact, I observed the exact opposite in the experiments: Using the "input worst" data sets delivers better results in outputs than "input best". Moreover, "output worst" delivers better results in inputs than "output best". This may just be a random

effect, but it clearly shows that good quality results for inputs do not necessarily indicate good quality results for outputs. Nevertheless, when I compare "input best/worst" with results for inputs and "output best/worst" with results for outputs, the order is as expected and the "best" upscalings deliver better results than the "worst" upscalings.

3.3.3 Spatial cluster distribution

To get an impression how these different upscaling methods partition the world spatially I have plotted the clusters for the case of 43 clusters derived from the original data set with grain size $0.5° \times 0.5°$ under use of all previously defined input data sets (Figure 3.6, 3.7, 3.8 and 3.9). I have chosen 43 clusters as it is the highest number of clusters which is still presentable graphically and has at the same time for MAgPIE a counterpart in terms of static grid upscaling (in this case with grain size of $60.0° \times 60.0°$).

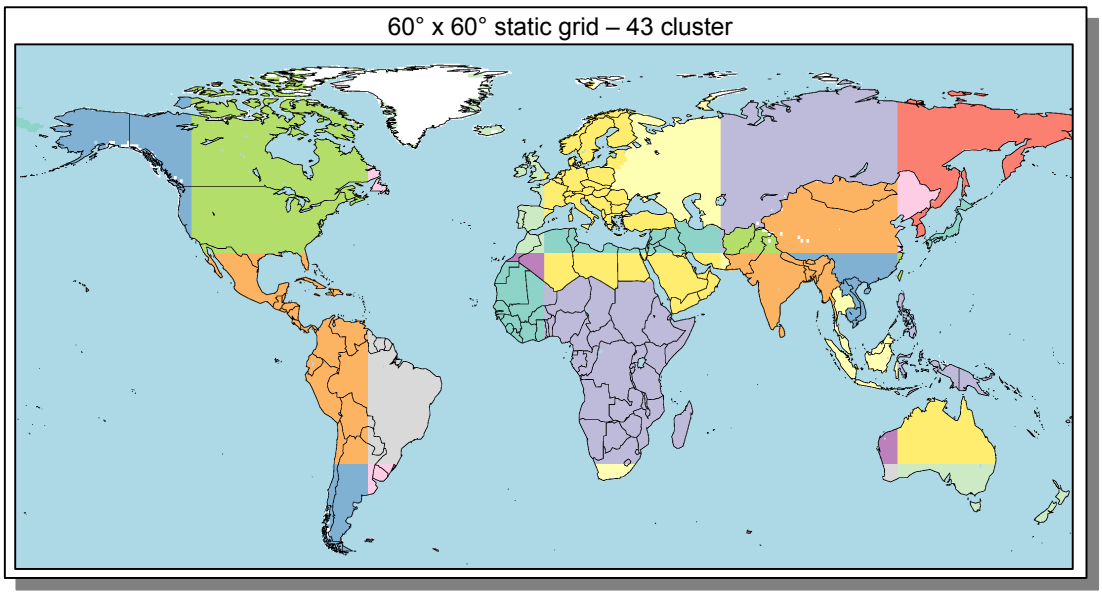

Figure 3.6: Map showing the 43 clusters used for a $60° \times 60°$ static grid upscaling (Different colors mean different cluster. 12 colors are used to distinguish between the 43 cluster. Since clusters belong to exactly one world region and none of the regions is containing more than 12 cluster, this distinction is unique.)

The map of clusters derived with the static grid upscaling methods shows clearly its geometric origin (Figure 3.6). The borders between the $60° \times 60°$ squares are easily to detect. However, at many locations the geometric structure is disrupted by the country-specific world region allocation of the 10 MAgPIE regions (compare Figure 2.2 and Appendix Table 2). This effect is caused by the MAgPIE requirement that any cluster has to belong exactly to one world region. Therefore the $60° \times 60°$ squares are split at region boundaries. Combined with the spatial structure of continents this leads to significant differences in cluster sizes. Some clusters are huge (e.g. the green cluster in

North America), other ones are tiny (e.g. the pink cluster in southern Latin America).

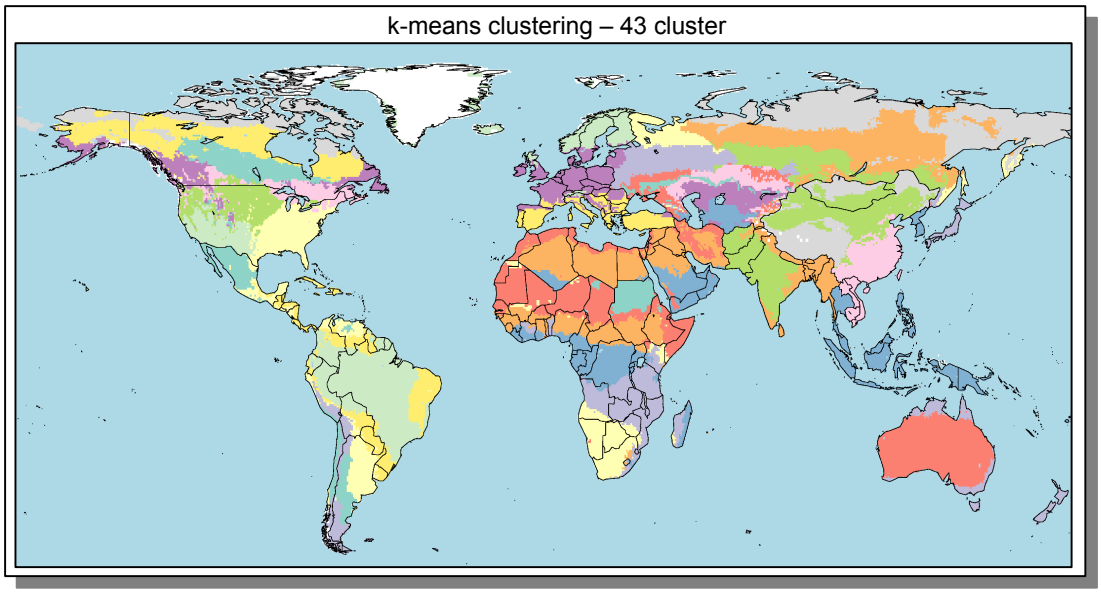

Figure 3.7: Map showing 43 clusters calculated with k-means clustering (Different colors mean different cluster. 12 colors are used to distinguish between the 43 cluster. Any cluster belongs exactly to one world region. No region contains more than 12 cluster.)

In contrast to static grid upscaling the clusters derived with the k-means method show its strong dependence on spatial, biophysical conditions, which were used as input for the clustering (Figure 3.7). Even though the location was not an explicit clustering criteria most clusters form a big, connected cluster core with some smaller, spatially detached parts. At many locations the clusters resemble well-known geographical structures such as deserts (Australia, North Africa), but also the dependence of many biophysical characteristics on the latitude becomes visible (longish clusters parallel to the equator such as in Canada, Russia or North Africa). Clusters are distributed relative homogeneously over the whole world.

As observed for k-means clustering also hierarchical bottom-up clustering resembles many geographical structures (Figure 3.8). Some clusters are nearly identical to the k-means results (especially for Australia and Europe), other ones differ significantly (Russia). In contrast to k-means the clusters are distributed less homogeneously over the world. Some regions, such as Pacific Asia, are clustered in more detail, other ones, such as Former Soviet Union, are represented with less cluster.

The map derived with top-down hierarchical clustering (Figure 3.9) shows many analogies to the result derived with bottom-up hierarchical clustering. Many structures are nearly identical, such as the huge cluster in Russia and Canada or the cluster covering Brazil. In most cases the boundaries between clusters are only slightly shifted. A systematic difference in both cluster maps is not visible.

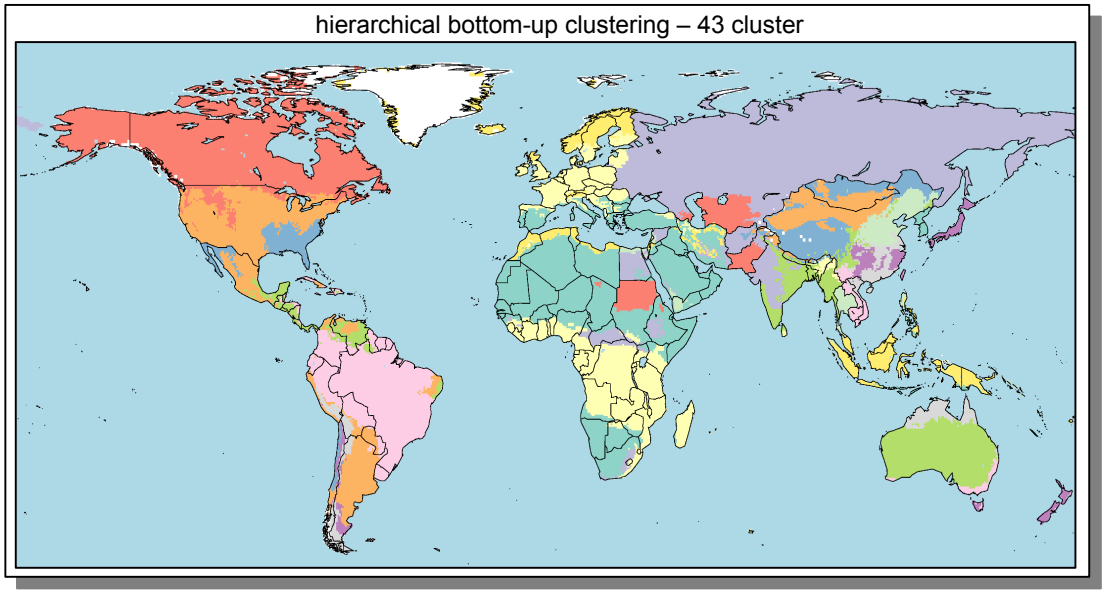

Figure 3.8: Map showing 43 clusters calculated with hierarchical bottom-up clustering (Different colors mean different cluster. 12 colors are used to distinguish between the 43 cluster. Any cluster belongs exactly to one world region. No region contains more than 12 cluster.)

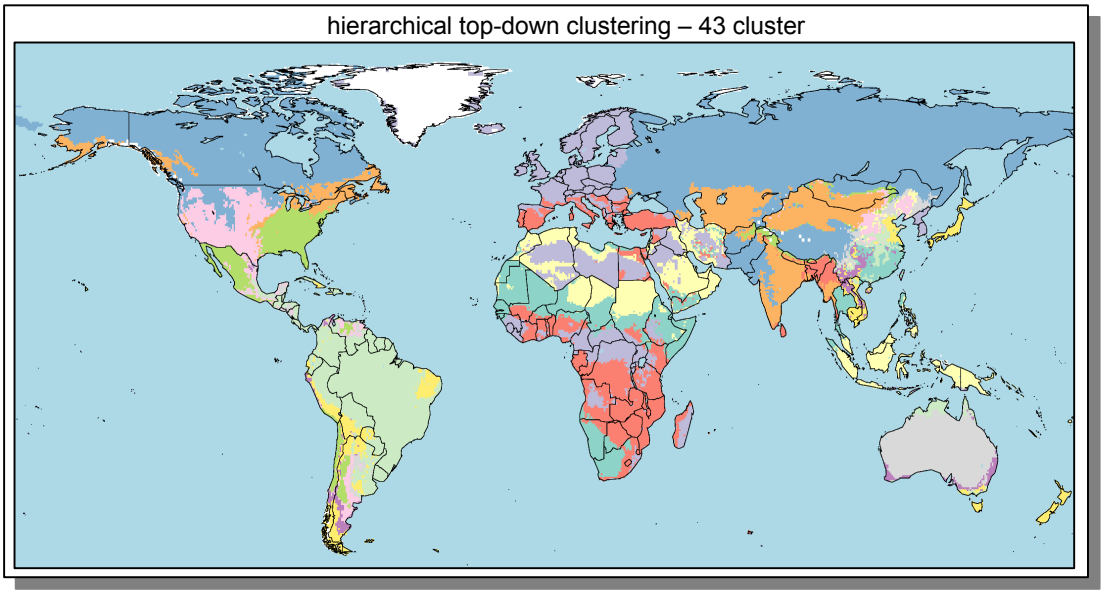

Figure 3.9: Map showing 43 clusters calculated with hierarchical top-down clustering (Different colors mean different cluster. 12 colors are used to distinguish between the 43 cluster. Any cluster belongs exactly to one world region. No region contains more than 12 cluster.)

3.4 Discussion

Agricultural land-use models combine processes across different scales. While some economic processes, like commodity trade, occur at the global scale, ecological parameters and farming decisions come into play at regional to local scales. Linking these scales is in land-use modeling as well as in many other research areas one of the important issues that have to be faced [Wessman, 1992, Cash and Moser, 2000, Harvey, 2000]. Upscaling through clustering is one option to improve these linkages and to enhance model precision.

The results show that all presented clustering methods deliver at least equal but mostly better results compared to standard upscaling based on static grids. Whereas the application of clustering methods compared to static grid upscaling is superior in all situations, the explicit choice of the clustering method is a matter of application and number of clusters. My investigations show that for comparing original and upscaled data sets directly two cases can be distinguished. Working with a few clusters (less than 500 clusters in the case of 2.0° grain size), k-means clustering delivers the lowest information losses and highest resolution conservation due to upscaling. Moving up to a higher number of clusters, hierarchical bottom-up clustering with maximal linkage becomes the best choice. Hierarchical top-down behaves over the whole range slightly worse than hierarchical bottom-up. This is in contradiction to my initial hypothesis that top-down clustering is the best choice for a few clusters whereas bottom-up works better for many clusters. It seems, that the advantage of top-down clustering not to have a significant lock-in effect for a few clusters is just dominated by other effects. Since the maximum linkage method only determines that the two cells with the highest distance within a cluster have to be separated in an iteration step, but does not make any decisions on the mapping of the remaining cells, there are several possible representations of maximum-linkage top-down clustering possible. So, the bad performance of top-down clustering may just be an effect of the chosen algorithm and not of the method itself. That hierarchical top-down clustering can be the best choice in some cases was shown by Steinbach et al. [2000] performing document clustering with a mixture of hierarchical top-down and k-means algorithm ("bisecting k-means").

Comparing outputs from model runs with original data and runs with upscaled data, the ranking between different clustering methods changes. The advantages of k-means clustering for a few cells compared to hierarchical bottom-up are not visible anymore and hierarchical bottom-up clustering shows now best results over the whole range. A possible explanation for this behavior is that k-means clustering is based on euclidean distances, whereas both hierarchical clustering methods use a maximum linkage approach which is a representation of the infinity norm. This leads to significant differences in the handling of outliers. In the case of the infinity norm outliers are preferably used as single clusters since only the maximum distance between two clusters is counted and not the sum of all connections between them. In the case of an euclidean distance outliers will be embedded much more often in an existing cluster since it not only accounts for the maximum distance but also for all distances between cells of both clusters. Looking at the upscaled data itself it is preferable to use a well-balanced distance measure, such as

the euclidean distance, since it accounts for any distance of any cell. Looking specifically at outputs of optimization models, extreme values as supplied by outliers are relevant for the optimization result. Hence, in order to rebuild the result produced with the reference data itself it is preferable to describe outliers as single clusters. This will reduce the loss of information which is most relevant for the modeling output. Consequently, the maximal linkage method is the best choice for preserving information most relevant for the optimization. Anyhow, when working with real data especially the correctness of outliers is often extremely arguable. From that point of view k-means clustering is reducing the similarity between outputs of reference data and upscaled data, but is probably reducing the biases caused by flawed outliers. Hence, the choice of clustering method is also a question of reliability of input data.

The role of outliers can also explain the fact that results for k-means are even worse for non-cellular outputs than for cellular outputs. Whereas in the case of cellular outputs still each single cell plays an important role for the total result, non-cellular outputs can be much more dominated by outlier-induced effects. But also for all other clustering methods results for non-cellular data are less significant than for cellular ones. This is also due to the fact that for non-cellular data errors from upscaling more frequently cancel each other out. This could explain the high randomness in the quality of results and also why even the reduction of clusters can improve results.

The comparison of different combinations of data sets to determine the clustering patterns shows that it is always a good choice to use any data set for clustering that also should be upscaled. A good selection of data sets can also deliver good clustering results, but cannot compete with results of an upscaling with all available data sets. Hence, for the clustering procedure the best choice is to use the full range of data sets that are going to be upscaled.

The comparison of cluster maps brings out the differences between the used methods graphically. The disadvantages of static grid aggregation become clearly visible: Geographical characteristics are not taken into account and region or continent borders lead to unwanted differences in cluster sizes. However, the latter effect is especially occurring for huge grain sizes and decreases with decreasing grain size. In contrast to static grid upscaling all clustering methods show clear responses to the spatially explicit, biophysical conditions. Several geographical structures can be found in a similar way in all cluster maps. While both hierarchical methods deliver nearly identical results, there is a significant difference between k-means and hierarchical clustering: Whereas k-means delivers a relatively homogeneous, spatial distribution of cluster, hierarchical clustering shows strong differences in cluster sizes. This is most obvious for Former Soviet Union, which is represented with only one cluster in both hierarchical maps, whereas k-means is using several clusters for it. This behavior has primarily two reasons: First, the maximum linkage method which was applied for hierarchical clustering favors the creating of very small cluster, while the k-means procedure minimizing the within-sum of squares leads to more homogeneously distributed cluster. Second, the number of clusters per region was predefined for k-means clustering (as the missing hierarchy does not allow for a reasonable adaption method). At the same time it was chosen dynamical for the hierarchical methods based on its hierarchy. Though, k-means was forced to use more

clusters for Former Soviet Union, while hierarchical clustering could shift these clusters to other regions.

3.5 Conclusion

In land use modeling, the combination of high-resolution data with high-complexity models remains a challenge. One approach to deal with it is clustering. My tests show that especially hierarchical bottom-up clustering leads to a significant reduction of information loss due to upscaling. Other clustering methods (e.g. hierarchical top-down or k-means clustering) also increase the quality of upscaled data, but to a lesser extend.

When using clustering in land use models, one also has to deal with the problem of interpolation. Before the actual clustering procedure it is important to determine the components of input data that are relevant for the cluster structure. One has to find interpolation rules for all other data sets of the model, since they cannot be assumed to be homogeneous within a cluster, if they were not part of the clustering.

Clustering alone does not solve the scale problem in land use models, but it allows - independent of the particular choice of a clustering method - to increase model accuracy and to reduce information losses within the upscaling process. Compared to spatial upscaling with static grids, clustering showed in my analysis always superior results and thus is the better choice for upscaling.

4 Measuring agricultural land-use intensity

Abstract

[1] Human activities such as research & development, infrastructure or management, are of major importance for agricultural productivity. Its amplitude can be captured with the concept of agricultural land-use intensity. I show, that agricultural land-use intensity can be seen as the human-induced amplification of yields. Furthermore, I present a measure, called the τ-factor, which is a complementary approach to current measures for agricultural land-use intensity. Whenever these current methods fail, the τ-factor becomes an alternative. The τ-factor is the ratio between actual yield and a yield under well defined management and technology conditions. By taking this ratio, the physical component, which is equal for both terms, is removed. This allows to analyze which regions have strongly amplified their physical yield levels and which have not. As an example I use this measure in combination with yields deduced from a global vegetation model for a global study of agricultural land-use intensities of the year 1995. Outputs are average land-use intensities in different world regions as well as land-use intensities for specific crops. The analysis shows that parts of North America, Russia and especially Africa had low agricultural land-use intensities, which implies good long-term potentials for further yield increases, whereas the Eastern US and Western Europe had already high agricultural land-use intensities in 1995.

4.1 Introduction

Together with expansion of the land area under agricultural production, agricultural land-use intensification is the major driver to satisfy future demand for agricultural products. Because land expansion is limited, intensification will become even more important in the future [Ewert et al., 2005]. To assess the long-term potential of further yield growth and to be able to make projections about future land-use developments it is essential to know current levels of agricultural land-use intensity. The ability to satisfy future demands and to prevent food crises but also the future development of food prices all depend on the potential for further land-use intensification.

Literature provides two different concepts for analyzing agricultural potentials: yield gap analysis and analysis of land-use intensities. Yield gap analysis is based on the concept that growth in agricultural production, like economic growth in general, can be attributed to two sources. First, it can be driven by growth in inputs (like labor, land or capital) and second, by gains in productivity or technological change respectively [Romer,

[1] This chapter is based on Dietrich et al. [2011c]

concept	description
yield	a measure of the output per unit area
human activity	any kind of human interaction influencing yields (e.g. management or R&D)
physical environment	natural circumstances under which production takes place (soil, climate, terrain)
agricultural land-use intensity α	degree of yield amplification caused by human activities
τ-factor	measure proportional to agricultural land-use intensity
agricultural long-term potential	potential for yield increases over the next decades based on current agricultural land-use intensities

Table 4.1: Concepts and terms used in this chapter

1990, Fulginiti et al., 2004]. It is assumed that each location on earth has an upper yield bound, called either "potential yield" [van Ittersum and Rabbinge, 1997] or "technology frontier" [Nishimizu and Page, 1982], which is set by present physical conditions and available technologies. Observed or actual yields may be lower than the potential yield due to the ineffective application of inputs and available technologies. Regions with strong discrepancies between actual and potential yield have a strong potential for further yield increases, whereas regions at the technology frontier are not able to increase their yields any further [Färe et al., 1994, Coelli and Rao, 2005, Neumann et al., 2010]. This approach can be seen as a short-term analysis of agricultural potentials, since it focuses on agricultural inputs and management, which can be changed and optimized within years, but excludes changes due to Research & Development (R&D), which typically have a time lag of around 10-30 years [Alston et al., 1998b, Alston, 2000]. For measuring distances of firms to the frontier, Shephard [1970] introduced the method of distance functions. Besides the economic performance, several studies include environmental indicators in this analysis [Färe et al., 1996, Reinhard et al., 1999, Munksgaard et al., 2007, Bellenger and Herlihy, 2009].

In contrast, the concept of agricultural land-use intensity does not measure the distance to a technology frontier or potential yield. Instead it is a productivity measure which is only taking the human-induced productivity into account, which - in contrast to yield gap analysis - also includes technological change as a source of growth. Although, both measures are calculated in some cases quite similar, their meaning is usually significantly different, as shown by the following thought experiment: Having a field with an actual yield equal to the potential yield the yield gap would be zero. Assuming now, that some technological progress takes place which is shifting the potential yield, but which is not adopted by the farmer (e.g. because the farmer does not get any information about the new technology) this would mean, that the yield gap increases. At the same time the land-use intensity remains constant, since the farmer did not improve his current management strategies but also did not downgrade it. So, in terms of yield gap one would observe a change for the worse, whereas the land use intensity measure would

report no changes at all. The yield gap measures the distance to the currently best, whereas the land-use intensity is measuring the absolute, human-induced productivity. This makes land-use intensity the more adequate tool for assessing long-term potentials and developments in agriculture (time horizons of several decades). Because of its broader spectrum of included activities it is also more appropriate as a surrogate for the general state of agricultural development. In a similar manner it is used in other fields as for instance in ecology, where several studies show a correlation between agricultural land-use intensities and decreases in species diversity [Oehl et al., 2003, Zechmeister and Moser, 2001].

The concept of land-use intensity is less clearly defined than that of yield gap analysis. In the literature one can find several measures for land-use intensity, but only a few are provided with an explicit definition of the underlying concept [Shriar, 2000]. I only found explicit definitions for the term "land-use intensification", but no one for "land-use intensity". Brookfield [1993] describes intensification as 'in relation to constant land, the substitution of labor, capital or technology for land, in any combination, so as to obtain higher long-term production from the same area'. Kates et al. [1993] and Netting [1993] use the formulation that intensification is 'a process of increasing the utilization or productivity of land currently under production, and it contrasts with expansion, that is, the extension of land under cultivation'. Shriar [2000] uses the formulation that 'agricultural intensification is a process of raising land productivity over time through increases in inputs of one form or another on a per unit area basis'.

Land-use intensity is measured either in an output-oriented or input-oriented way using inputs as surrogates for increases in productivity [Lambin et al., 2000]. When the focus is on output, intensity can be measured in production units (calories, tons, monetary value,...) per area per time unit [Turner and Doolittle, 1978]. In an input-oriented approach the amount of inputs is measured and weighted with their assumed increase in production [Shriar, 2000, Turner and Doolittle, 1978] or single input characteristics are used as a surrogate, for instance cultivation frequency [Boserup, 2005].

Measuring intensity using output should be the most appropriate way but it suffers several disadvantages: depending on the region under investigation and the intended resolution and accuracy, yield data may be unavailable, whereas data on applied inputs may exist. Comparing different crops concerning their land-use intensity is problematic because of differing yield levels, which also depend on the chosen production unit. Hence, one needs a common denominator to be able to compare land-use intensities of different crops against each other [Kates et al., 1993]. Using inputs as surrogates for land-use intensity has in some cases the advantage of better data availability, which can also make it the only option [Shriar, 2000]. However, this approach requires detailed assumptions about the kind of inputs and their contribution to total output. One also needs to know exactly which inputs are relevant and have to be taken into account.

Comparing all these definitions and measures one can find general agreement in two areas: (1) intensification means increases in productivity and (2) intensification can be achieved by a broad spectrum of options which are all induced by humans. Changes in productivity due to environmental reasons are excluded. Based on this consensus I present a slightly different definition of intensification: agricultural land-use intensifica-

tion is the increase of land productivity due to human activities. And in line with this definition agricultural land-use intensity is defined as the degree of yield amplification caused by human activities. These definitions are quite similar to former ones but highlight that any kind of human interaction with agriculture that affects productivity also affects land-use intensity whereas no kind of environmental interaction has any influence on it.

In order to assess the long-term potential for further yield growth one needs to know current agricultural land-use intensities. In a region with high agricultural land-use intensity many known options to increase yields are already being applied. This does not inhibit further intensification, but raises the marginal effort that is required for further enhancements [Jones, 2009]. In other words, a high agricultural land-use intensity means higher marginal costs for further yield increases, which makes investments in this region less attractive.

To capture agricultural land-use intensities I present a mathematical description based on the formal definition. Based on this description I derive another measure, the "τ-factor", which is easier to calculate than the land-use intensity itself but differs only in a scalar value. It is the ratio between an actual yield and a yield that would be achieved under a constant land-use intensity. It is defined in a general way but for the calculation it is necessary to have data on actual yields and yields under a known land-use intensity. That data can either be derived by models or statistical analysis. For the exemplary application of the methodology I used yield data simulated by a global vegetation-crop model, the "Lund-Potsdam-Jena dynamic global vegetation model with managed Land" (LPJmL) [Bondeau et al., 2007].

Compared to other surrogate measures for land-use intensity, which are typically input-oriented, such as the mentioned cultivation frequency [Boserup, 2005] or other input oriented approaches [Shriar, 2000, Turner and Doolittle, 1978], the τ-factor exploits the contrary, output-oriented approach. Its complementarity makes it a beneficial alternative whenever current methods for estimation of land-use intensity fail. For instance this is the case, if output data has a better availability than data on inputs, if the exact contribution of inputs to agricultural productivity is unclear, or if the set of inputs relevant for agricultural productivity cannot be fully recovered. Performing a global analysis, all these three mentioned conditions are part of the general issue, which makes the τ-factor the preferable choice for this task.

This chapter is structured as follows: Section 4.2.1 describes the basic methodology, section 4.2.2 explains the utilization of modeled crop yields in computing τ and section 4.2.3 deals with the spatial aggregation of the τ-factor. These parts are followed by a presentation of results of a global analysis (Section 4.3) and a discussion of results and methodological aspects (Section 4.4). The chapter concludes (Section 4.5) with a recapitulation of the approach and its applicability in future studies.

4.2 Methods

4.2.1 Theoretical framework - agricultural land-use intensity and τ-factor

Crop yields are useful parameters in assessing agricultural land-use intensity. The yield of a region itself already provides a rough estimate of it. However, this measure is still distorted because of its dependence on the physical environment. Hence, a high yield could either indicate a high agricultural land-use intensity or favorable physical conditions.

$$Y(c,j) = \underbrace{\alpha(c,j)}_{\text{land-use intensity}} \cdot \underbrace{Y_0(c,j)}_{\text{base yield}} \tag{4.1}$$

Applying the concept of partial factor productivity [Nin et al., 2003], a yield $Y(c,j)$ (of crop c at place j) can be described as the product of two factors: a base yield $Y_0(c,j)$ and an amplification factor $\alpha(c,j)$ (Equation 4.1). The base yield depends only on the physical environment and is free of any human influences except sowing and harvesting, which are essential cropping activities. In contrast, the amplification factor α is independent of the physical environment and represents only the amplification of yields due to human activities. Thus α is the agricultural land-use intensity.

The base yield Y_0 is a highly theoretical construct. Typically there is no clear border between the physical environment and human activity. Hence, it is problematic to calculate the base yield Y_0 itself. As a workaround I use a reference yield Y_{ref} which has only to fulfill the requirement that its agricultural land-use intensity α_{ref} is constant for all crops and at any place. Dividing the actual yield Y_{act} by this reference yield Y_{ref} (Equation 4.2) results in the τ-factor. It is independent of the physical environment, since Y_0 remains the same for both yields. Furthermore, it is proportional to the agricultural land-use intensity α_{act} but easier to calculate since the only requirement for Y_{ref} is an equal agricultural land-use intensity at any place and for any crop. Y_{ref} can be modeled much more straightforward compared to Y_0 which has the much stronger requirement of total absence of agricultural land-use intensification. The τ-factor is not the land-use intensity itself but can be used as a surrogate, as it is only a scaled version of it. The scaling only becomes important when comparing differently calculated τ-factors with each other.

$$\tau(c,j) = \frac{Y_{act}(c,j)}{Y_{ref}(c,j)} = \frac{\alpha_{act}(c,j) \cdot Y_0(c,j)}{\alpha_{ref} \cdot Y_0(c,j)} = \frac{\alpha_{act}(c,j)}{\alpha_{ref}} \tag{4.2}$$

In general the actual yield can be measured directly, whereas the reference yield Y_{ref} has to be deduced theoretically. The τ-factor has a close relationship to the yield definition. It displays the same responsiveness to changes in management, technological progress and other human activities as crop yields: a general yield growth driven by intensification results in the same growth of yields and τ assuming that physical conditions have not changed. A τ-factor in one region A that is twice as high as a τ-factor in another region B can be interpreted in the following way: if both regions have the

same physical conditions region A will have twice the yield of region B due to a higher agricultural land-use intensity. If, on the contrary, physical conditions in region A are half as good as in region B, yields in region A would equal yields in region B if τ_A is twice as high as τ_B. Thus, the τ-factor not only ranks regions, but also delivers quantitative information about yield differences between regions due to differences in human activity (agricultural land-use intensity).

4.2.2 Calculating the τ-factor

Before calculating τ it is important to know how this factor works. Roughly speaking, the factor compares a yield simulated by a model that captures some, but not all aspects of agriculture, with observations. The τ-factor measures exactly the fraction that is not part of the model. This approach has the advantage that one does not need to know what human activities are contributing to agricultural land-use intensification. Therefore, the approach can also capture impacts of activities that are not yet known to be relevant for agriculture. On the other hand, it has the disadvantage that anything else that is not part of the model is captured as well. Hence, one needs a model which is accurate in simulating those parts which are to be excluded from the measurement - in this case the physical environment. This is done best by a vegetation model without explicit implementation of agricultural management (or at least with the option to switch off management).

Another problem of this analysis is the consistency of yield data. Actual yield data is available at national [FAOSTAT, 2009] or even sub-national level [Monfreda et al., 2008] and reference yield data has to be deduced theoretically, e.g. by means of crop models. However, especially global models typically suffer from systematic errors and biases caused by the high complexity of the underlying problem. To reduce the error caused by these biases in τ, consistency between actual and reference yields plays a major role. Therefore, the simulation of reference yields and the model-based downscaling of FAOSTAT data as actual yields is done with the same model. By doing so, impacts of model biases on τ do not vanish but are reduced, since they appear in the same way in numerator and denominator.

In this dissertation I use the "Lund-Potsdam-Jena dynamic global vegetation model with managed Land" (LPJmL) which simulates these yields at sub-country level ($0.5° \times 0.5°$ grain size) (Chapter 2.4). The used LPJmL version provided yields for 11 crop-types[2]. This model choice, its special characteristics and the corresponding results should only be seen as an exemplary application of the presented method. The quality of the results significantly depends on the quality and specialization of the model. Furthermore, the concrete procedure to calculate reference yields and to downscale actual yields also strongly depends on the chosen model.

LPJmL represents the human impact on agriculture via the maximum Leaf Area Index (LAImax), an index that depicts the ratio of leaf surface to covered land surface and affects the overall productivity of the plant via the fraction of absorbed photosyn-

[2]wheat, rice, maize, millet, pulses, sugar beet, cassava, sunflower, soybean, groundnut and rapeseed

thetically active radiation (fpar), the scaling factor from leaf-level photosynthesis to field scale (alphaa, representing the homogeneity of a field), as well as the harvest index (HI), assuming that intensive systems grow high yielding varieties and extensive systems grow more robust but lower-yielding varieties [Gosme et al., 2010]. All three factors are directly linked: highly developed systems are parameterized with a high LAImax value, high alphaa, and high HI [Fader et al., 2010]. The value of LAImax varies in integer steps from 1 to 7 which is in agreement with the observed range of values. A high agricultural land-use intensity is connected to a high LAImax which in turn produces high yields. Setting LAImax fixed to the same value globally leads to simulated yields that represent homogeneous agricultural land-use intensities.

For downscaling of actual yields to sub-country level all 7 LAImax steps were simulated and compared with observed data from FAOSTAT [FAOSTAT, 2009]. For each crop in each country that LAImax-value was chosen which best reproduced observed national yield. This heterogeneous country- and crop-specific LAImax map was then used in LPJmL for calculating actual yields Y_{act} at sub-country level. Important in this context is that these 7 steps cover a huge range of yields and therefore allow one to reproduce any yield that is more or less realistic. So the actual yields primarily have to be seen as a model-assisted downscaling of FAO data and only to a lesser extent as a modeling result. The reproduction of observed yields via the model also increases the consistency of the input data, as reference yield and actual yield are thus directly comparable.

Before simulating the reference yields Y_{ref} one has to decide what the reference land-use intensity α_{ref} should look like. Methodologically, this choice does not affect the results, and hence one could just take in the example one of the seven runs with fixed LAImax. However, to reduce the impact of model-based errors on the result and to increase the signal-to-noise ratio of the simulated data, I decided to take the mean of all seven LAImax simulations as the reference yield. Hence the reference land-use intensity α_{ref} is equal to the mean of all seven agricultural land-use intensities (Equation 4.3).

$$\alpha_{ref} = \frac{\alpha_{LAImax=1} + \alpha_{LAImax=2} + ... + \alpha_{LAImax=6} + \alpha_{LAImax=7}}{7} \tag{4.3}$$

4.2.3 Aggregating the τ-factor

Based on the yield data one is able to calculate τ for each grid cell. However, in several cases it is more useful to have one mean value for a whole region instead of one value per cell. This would, for example, have the advantage of a better signal-to-noise ratio because cell-specific errors are reduced. In addition, this aggregation may be necessary for specific research questions, e.g. when investigating relationships between investments in Research & Development and the τ-factor (Chapter 5): to reduce the influence of spillovers, τ-values have to be aggregated to a level higher than the average spillover range.

Starting with equation 4.2 the aggregated τ-factor can be written as the ratio of

aggregated actual yield Y_{act} and aggregated reference yield Y_{ref} (Equation 4.4).

$$\overline{\tau}(c,i) = \frac{\overline{Y_{act}}(c,j)}{\overline{Y_{ref}}(c,j)} = \frac{\sum_{j \in i} \overbrace{\frac{Y_{act}(c,j)}{Y_{ref}(c,j)}}^{\tau(c,j)} \cdot \overbrace{Y_{ref}(c,j) \cdot A(c,j)}^{X_{ref}(c,j)}}{\sum_{j \in i} \underbrace{Y_{ref}(c,j) \cdot A(c,j)}_{X_{ref}(c,j)}} \tag{4.4}$$

After expanding the crop-area-weighted yield aggregation (with crop area $A(c,j)$) and some transformations one gets an equation for a weighted τ-factor aggregation (Equation 4.5) with weight X_{ref}.

$$\overline{\tau}(c,i) = \frac{\sum_{j \in i} \tau(c,j) \cdot X_{ref}(c,j)}{\sum_{j \in i} X_{ref}(c,j)} \tag{4.5}$$

The weight X_{ref} is a product of crop area $A(c,j)$ and reference yield $Y_{ref}(c,j)$ (Equation 4.6). It can be interpreted as total production under reference conditions.

$$X_{ref}(c,j) = Y_{ref}(c,j) \cdot A(c,j) \tag{4.6}$$

4.3 Results

4.3.1 τ-factor estimation

As an example for the general τ-factor estimation procedure I have plotted the global distributions of reference yield and actual yield for maize (Figure 4.1 and 4.2) and the resulting τ-factor distribution (Figure 4.3).

The reference yields (Figure 4.1) show a relatively homogeneous picture with maize yields mostly between 0 and 300 gC/m^2 (C = Carbon). Highest values can be found in parts of south and central Asia (Afghanistan, Kasachstan, Turkmenistan, Usbekistan) in the Nil river basin in Egypt, southern parts of Brazil, Chile, in parts of Australia and at the U.S. west coast. Low values can be found especially at high latitudes in Canada and Russia, but of course also in dry regions such as parts of Africa.

Looking at the actual yields for maize one finds a significantly higher variance in yields (Figure 4.2). Yields range from 0 to 550 gC/m^2. In some cases quite good yield potentials are used to receive high yield levels, as for example at the U.S. west coast, Chile or in the Nil river basin in Egypt. In other cases these potentials are not exploited, such as in South Brazil and Afghanistan. It is also interesting to note that some regions are using only average reference yields to receive quite high actual yields, such as parts of North America, Spain or Iran (last one is using lower reference yields compared to its direct neighbor Afghanistan to receive higher actual yields). Whereas for reference yields country-specific differences did not exist, they become apparent for actual yields (compare for instance Iran with Afghanistan, or Argentina with Brazil).

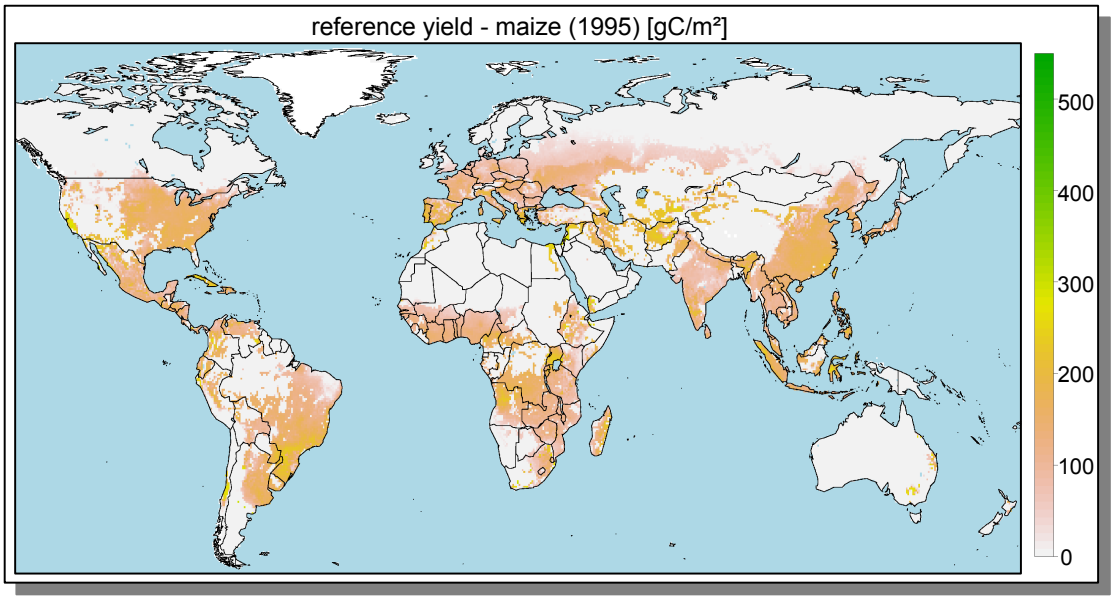

Figure 4.1: Global reference yield distribution for maize in 1995

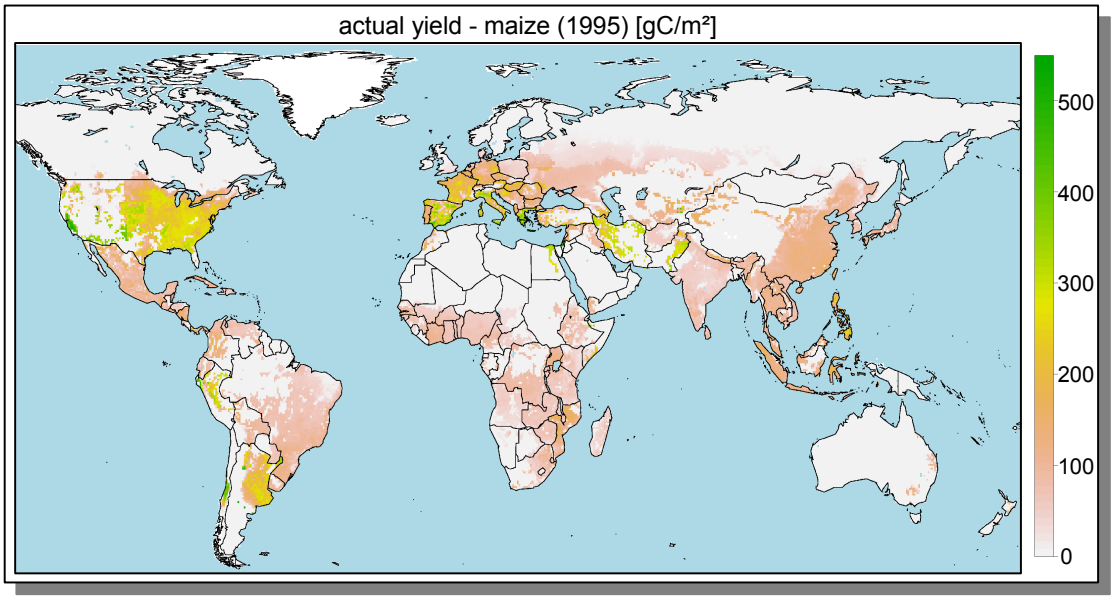

Figure 4.2: Global actual yield distribution for maize in 1995

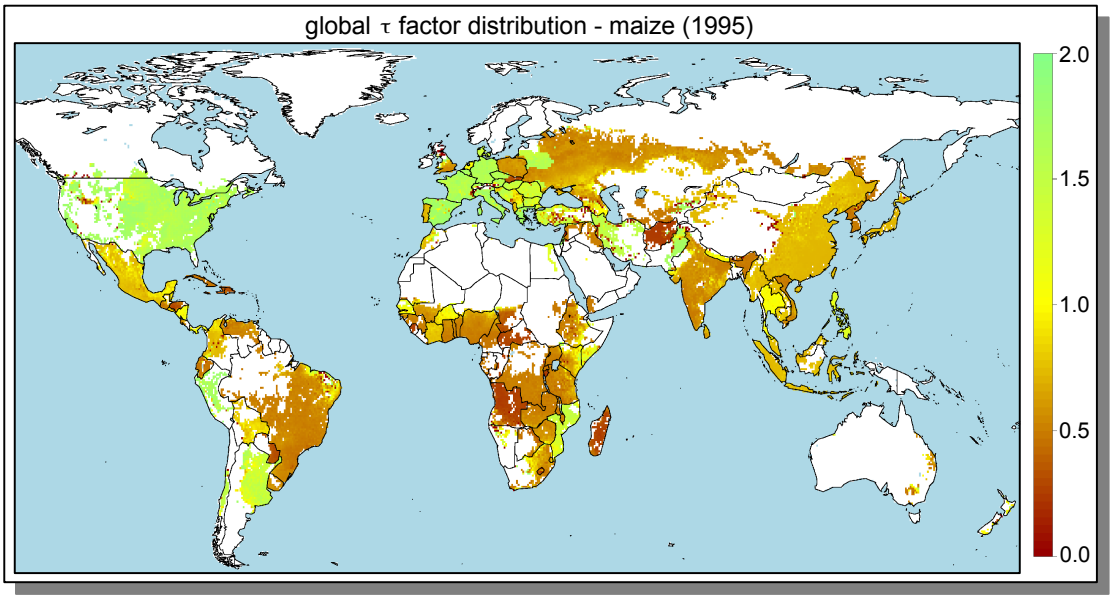

Figure 4.3: Global τ-factor distribution for maize in 1995

Taking the ratio between actual and reference yields one gets the τ-factor distribution (Figure 4.3). The trends already discussed for the comparison between reference and actual yields become now directly visible. For instance, the good performance of Iran and the bad performance of Afghanistan is easily to detect as well as the bad performance of Brazil and quite good performances of Argentina, Chile, Peru. As I have chosen an average yield level as reference yield the values vary between 0 and 2. As the actual yields are calibrated based on FAO country-data the borders between countries become now also prominently visible in the data. Furthermore, the variance within countries is reduced compared to the variance in yields within a country (good to see for instance for the U.S.). Most countries show quite homogeneous τ-values, often combined with some outliers. Outliers typically occur for cells with quite low yield levels, since small errors can have there a huge influence on the general result. Overall one gets a quite clear picture of countries with high and countries with low land-use intensities.

4.3.2 crop-unspecific τ-factor

To get a crop-unspecific, general τ-factor for each location one has to aggregate the values for the different crop-types. Figure 4.4 shows this procedure for North America (NAM) and the 11 crop-types supplied by the used LPJmL version. As one can see crop-specific τ-values within a country are relatively homogeneous, caused by the country-based calibration of actual yields. Inhomogeneities within countries are primarily caused by two factors: First, outliers due to low yield levels in reference and actual yields, so that small simulation errors have a huge impact on the results. Second, general broader scale variations, which are caused by slightly different responses of the yield levels to the

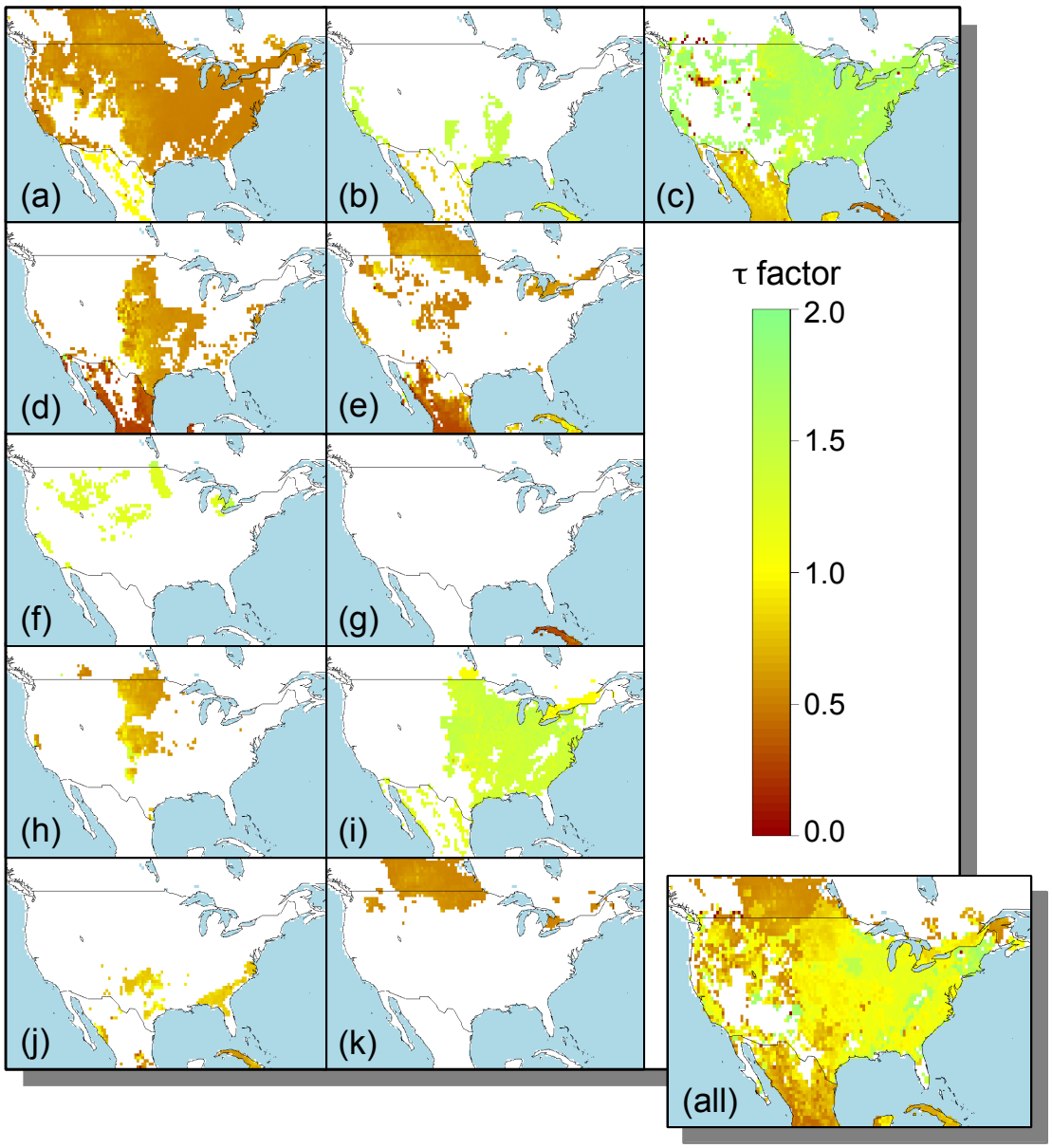

Figure 4.4: crop-specific τ-factors for North America (NAM) in 1995 and their aggregate. The corresponding crop-types are: (a) wheat, (b) rice, (c) maize, (d) millet, (e) pulses, (f) sugar beet, (g) cassava, (h) sunflower, (i) soybean, (j) groundnut, (k) rapeseed. The aggregated τ-factor, which is derived by calculating the mean over all crop-types for each cell, is marked as (all).

LPJmL management factors. However, the overall results are primarily influenced by the country-specific calibration to FAO yield levels. Aggregating these crop-specific values to the general τ-factor one gets a picture with higher variance within a country. In this case,

the variance is caused by the crop-specific differences in τ-factors in combination with the underlying land-use patterns describing the spatial crop distributions. Therefore, this combination of different crop-specific τ-factors, which were calibrated with country-specific data and the cellular land-use patterns allow to make statements concerning sub-national land-use intensity patterns. However, its resolution and accuracy is quite limited. In the presented example of North America the combination of different crop-specific τ-factors to a general τ-factor uncovers an East-West-divide with high land-use intensities in Eastern U.S. and lower intensities in Western U.S.. Furthermore, the combination of crop-specific τ-factors leads to smoother transitions between countries compared to intensities of single crops. For instance, differences in intensities between Mexico and the U.S. are quite obvious for wheat, rice, maize and millet. However, the transition for the generalized land-use intensity is relatively smooth at the border of both countries. Same holds true for the border between the U.S. and Canada. Whether this behavior has a content-related reason or if the higher aggregation level is just averaging errors in the calibration process out, is still unclear.

4.3.3 further results

crop types	**tece**	**maize**	**trce**	**rice**	**oilcrops**
AFR	0.61	0.55	0.60	0.77	0.58
CPA	0.51	0.73	0.51	1.23	0.73
EUR	1.01	1.45	1.22	1.64	1.08
FSU	0.60	0.67	0.56	1.16	0.81
LAM	0.76	0.76	0.67	1.06	1.00
MEA	0.53	1.11	0.65	1.54	1.01
NAM	0.53	1.64	0.59	1.46	1.03
PAO	0.64	0.67	1.06	1.70	0.67
PAS	0.76	0.85	0.75	1.13	0.82
SAS	0.49	0.63	0.58	1.18	0.66

Table 4.2: Crop-specific τ-factors in world regions (1995)

I have calculated τ-factors for 10 world regions[3] (see Table 2 in Appendix for country-to-region mapping) and five different commodity types (temperate cereals (tece), maize, tropical cereals (trce), rice and oil crops) for 1995 (Table 4.2). EUR has the highest τ-factors for oil crops, temperate and tropical cereals, NAM shows the highest maize τ-factor and PAO the highest τ-factor for rice. AFR has the lowest τ-factors for maize, rice and oil crops and CPA for temperate cereals and tropical cereals. Some regions have only slight variations in their τ-factors over all crops, e.g. AFR (constantly low) or EUR

[3]AFR = Sub-Sahara Africa, CPA = Centrally Planned Asia (incl. China), EUR = Europe (incl. Turkey), FSU = Former Soviet Union, LAM = Latin America, MEA = Middle East and North Africa, NAM = North America, PAO = Pacific OECD (Australia, Japan and New Zealand), PAS = Pacific Asia, SAS = South Asia (incl. India)

(constantly high), whereas other regions show strong variations between crops, as e.g. PAO or MEA.

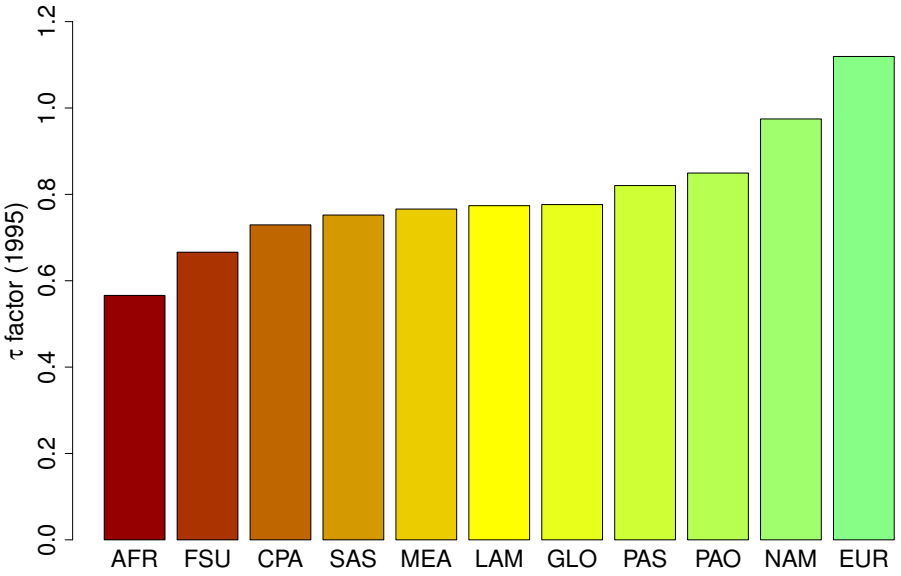

Figure 4.5: τ-factors in 1995 world regions & global (GLO)

Figure 4.5 shows the aggregated τ-factors over all crops for all regions and at global level in 1995 (crop area weighted mean using observed cropland shares reported by Portmann et al. [2010] & Fader et al. [2010]). The range is between 0.56 and 1.12. Comparing e.g. Sub-Saharan Africa (AFR) and Europe (EUR), the same physical conditions lead to a 98% higher yield in Europe compared to Sub-Saharan Africa due to its higher agricultural land-use intensity. Besides AFR which displays constantly low levels over all crops and EUR with constantly high levels, the general ranking shows a broad spectrum of regions close to the global mean of 0.78. Especially SAS, MEA and LAM are within a range from 0.75 to 0.77.

Figure 4.6 shows the global distribution of the mean τ-factors in 1995. Similar to the crop-specific τ-factors, one can observe regions with homogeneous spatial distribution of τ as well as regions with strong heterogeneity. For instance, the Republic of Ireland, France, Germany, Finland and Romania show relatively homogeneous values at a high level and Madagascar and Angola homogeneous values at a low level, whereas the US, England and South Africa seem to be more heterogeneous in their τ-factors. One finds high values in the Eastern US and Central Europe, medium values in Latin America and Asia and low values in Central Africa and Eastern Europe / Russia.

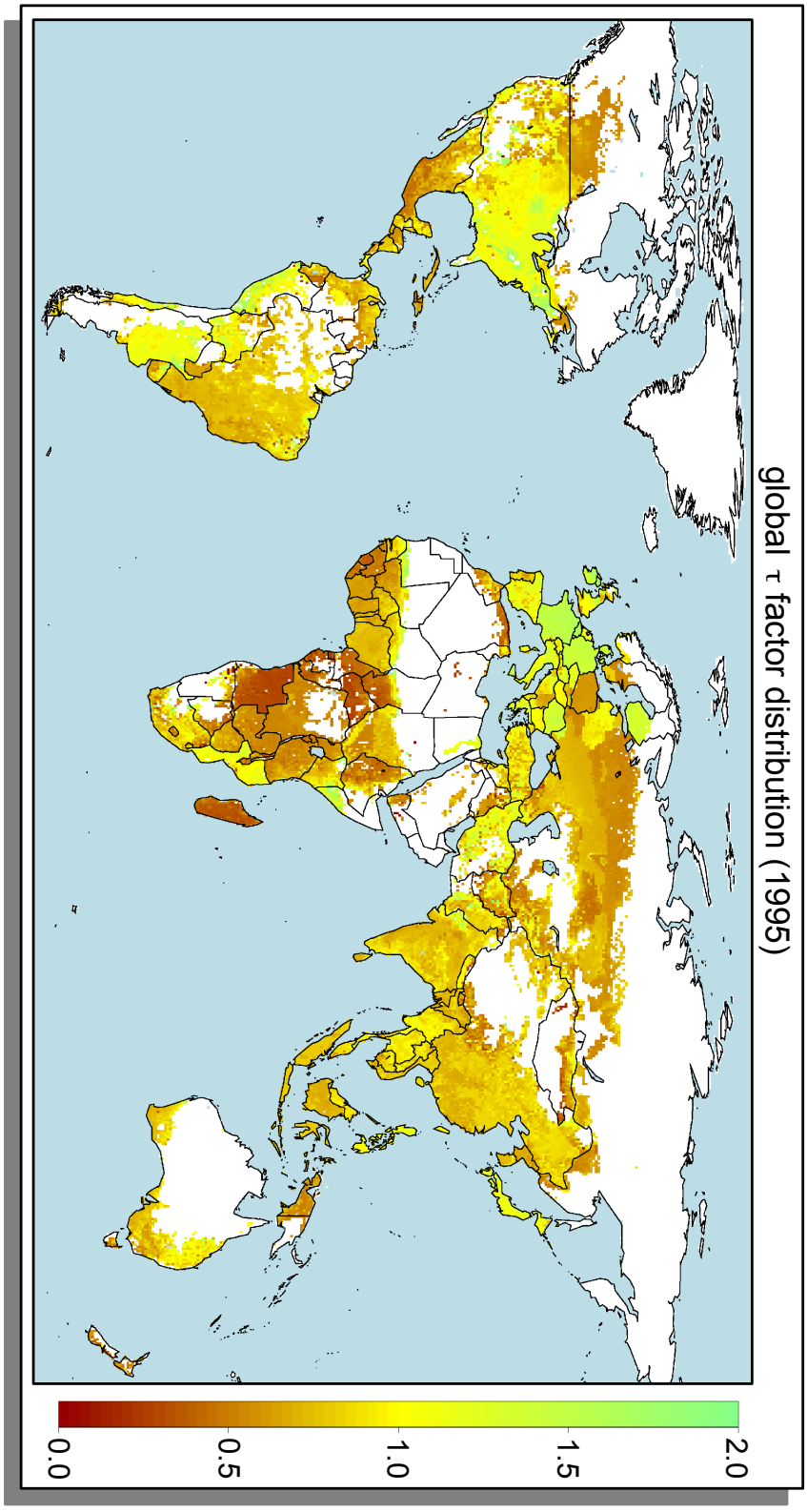

Figure 4.6: Global τ-factor distribution in 1995

4.4 Discussion

Methodological, the τ-factor is an extension to currently available measures for land-use intensity, with some advantages but also disadvantages in a direct comparison. Its most important difference is, that the τ-factor measures land-use intensity in an output-oriented manner, whereas currently applied methods typically use an input-oriented approach. Recent methods are the best choice, if detailed data on inputs exists and the relevant inputs and their impact on the total production is well known. If this is not the case, the τ-factor might be the better choice. However, to compute the τ-factor it is necessary to have detailed data on achieved yields as well as data about the relevant environmental factors. The different approaches also cause differences in the kind of result that is derived: Input-oriented measures for land-use intensity deliver, besides a general estimate of the total land-use intensity, also information about the relevance of certain inputs and the land-use intensity can be broken down to these inputs. However, the estimates of total land-use intensity are not complete, since they only take the defined inputs into account. The τ-factor delivers an estimate of the total land-use intensity which cannot be broken down. Though, the measure is complete in a sense that all impacts on agricultural intensity are taking into account, because of its output-related nature.

The quality of the calculated τ-factors strongly depends on the quality of the used models or statistical approaches to estimate reference yields. While this approach does not require detailed knowledge about the underlying human interactions it requires detailed knowledge on the environmental interactions. Having better data and knowledge on the environmental side, the τ-factor becomes the best choice; having a better base on the human side, input-based measures perform better.

Concerning the interpretation of the global analysis, it is important to realize that low agricultural land-use intensities do not directly imply strong yield increases in the future. There are many other factors that have to be taken into account when predicting future yield growth. For instance, Africa shows low land-use intensities, but weak institutions and political conditions in many African countries are elements that could inhibit its potential. Another factor is production costs which may increase disproportionately with yield. In this case it becomes uneconomical to produce at a higher land-use intensity. Typical examples for this behavior are sparsely populated regions with high wage levels, as for instance the Western US or Canada, where labor becomes a more limiting and more cost-determining factor than land [Runge et al., 2003, Federico, 2005].

Since this study is the first one assessing land-use intensities on a global level, it was problematic to find comparable studies. However, what I have found, is in a good agreement to this study: Compared to a global yield-gap analysis by Neumann et al. [2010] most regions show similar behavior in land-use intensities and efficiencies for maize. Exceptions are China and Brazil, for which efficiency levels are slightly higher compared to land-use intensity levels and Eastern Europe and Mexico, which are slightly less efficient than the land-use intensity suggests. Taking into account that yield-gap and land-use intensity are measuring slightly different things, this result is quite exciting: It shows, that in most regions technological development and management efficiency seem

to evolve identical. However, the exceptions show also that this does not have to be the case. A possible explanation for the differences is that Eastern Europe and Mexico have strong neighbors with Western Europe and USA, which China and Brazil do not have. So the yield gaps for Eastern Europe and Mexico are relatively high, because their direct neighbors show significantly better performances. On the other hand China and Brazil perform quite well in comparison to their direct neighbors, but are on an average level in a global comparison.

Comparing the presented results with FAO/OECD yield growth projections [OECD-FAO, 2009, Bruinsma, 2003] the general results are coherent: regions with high agricultural land-use intensities, for instance EUR or NAM, have low yield growth projections, whereas AFR has low land-use intensities and high yield growth projections. However, there are more differences between projections and intensities. For instance, FAO yield growth projections for SAS are significantly higher than for CPA, whereas land-use intensities, derived in this study, are slightly higher in SAS. This strengthens the perception that neither efficiencies derived with yield-gap analysis nor land-use intensities can be used in isolation to make yield growth predictions. There are several other aspects like demand and political conditions that influence yield growth. However, land-use intensities are still useful to obtain a coarse ranking concerning future yield growth. The τ-factor approach shows that Canada and the US still have significant long-term potential for cost-efficient yield increases whereas e.g. Western Europe is already at a high level of exploitation. Overall the highest long-term potentials for yield increases can be found in Africa, parts of North America and Eastern Europe / Russia. However, parts of Asia and Latin America (especially Brazil) also show a good base for further yield improvements.

4.5 Conclusion

The future development of agriculture is closely connected to current agricultural land-use intensities. For this study I have developed a output-based land-use intensity measure. One application for which this measure fits quite well is a global land-use intensity analysis. The analysis shows that Europe, North America and parts of Asia exhibit high agricultural land-use intensities whereas Africa and countries belonging to the Former Soviet Union display significantly lower land-use intensities. Concerning further yield increases one observes that the Western US, Canada and nearly the whole continent of Africa show high long-term growth potentials, whereas the Eastern US and Western Europe are already on a high level.

The τ-factor is a good alternative, whenever other methods fail, because of its complementary nature. Its replacement of the need for detailed socio-economic data by a requirement for data on the natural system makes it favorable for global studies. Furthermore, it is a useful measure for implementation of technological change and related R&D investments in land-use models as presented in the following chapter (Chapter 5).

5 Technological change in a global land-use model

Abstract

[1] Technological change in agriculture plays a decisive role for meeting future demands for agricultural goods. Especially in the longer run, i.e. several decades, technological change will be one of the major determinants of agricultural production. However, up to now, most agricultural sector models and models on land use change have used technological change as an exogenous input due to various information and data deficiencies. This chapter provides a first attempt towards an endogenous implementation based on the τ-factor. Empirical data on investments in technological change as well as production costs are correlated with the τ-factor. The estimated yield elasticity with respect to research investments in the presented approach is 0.27 and production costs per area increase linearly with an increasing yield level. Having implemented this approach in the global land-use model MAgPIE allows to make projections about yield growth rates in the future. Highest future yield increases are obtained for Sub-Saharan Africa, the Middle East, South and Pacific Asia. A validation with FAO data for the period 1960-2005 shows that the model behavior is in line with recent observations.

5.1 Introduction

More than 200 years ago Thomas Malthus published his rather pessimistic population essay, in which he was stating that population growth would be restricted by a low growth rate of food production [Malthus, 1998]. Now the world is inhabited by almost seven billion people, which marks an increase by about 600% since Malthus' times. One of the main shortcomings of his essay was the underestimation of technological change (TC - as defined in Table 5.1) in agriculture [Trewavas, 2002].

However, during Malthus' times technological change was negligible and higher food production was almost exclusively due to an increase in production factors [Federico, 2005]. Important innovations in agriculture from the 19th century onwards changed this pathway [Runge et al., 2003]. Since then land-saving technological change has been the main driver for growth in agricultural output [Wik et al., 2008, van Meijl and van Tongeren, 1999, Rosenzweig et al., 1988]. Figure 5.1 shows the strong correlation between agricultural output and population during the last 200 years. Agricultural

[1] This chapter is based on Dietrich et al. [2011b], which I have written in close collaboration with Christoph Schmitz, who has contributed to the same degree to the paper as I did.

concept	description
agricultural land-use intensity	degree of yield amplification caused by human activities
τ-factor	measure proportional to agricultural land-use intensity
technological change (TC)	more efficient usage of the input factors land, labor or capital [Romer, 1990]
TC investments	composite of annual investments in R&D and infrastructure (e.g. transport and telecommunication) [US$/year]
investment-yield ratio (IY ratio)	TC investments required per human-induced unit yield growth and area [US$/ha]

Table 5.1: Concepts and terms used in this chapter

output has increased considerably, paving the way for strong population growth. Most of such increases in agricultural output have been the result of technological change induced by investments in Research & Development (R&D). One example is the so called "Green Revolution" in Asia and Latin America, initiated by two international agricultural research institutes [Evenson and Gollin, 2003][2].

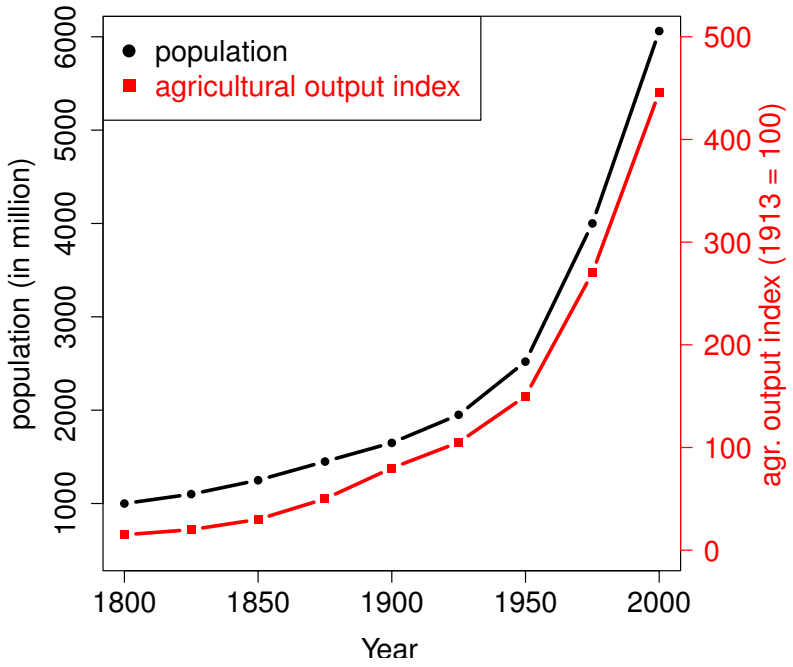

Figure 5.1: Historic development of agricultural production and population [own illustration based on Federico [2005] and United Nations [2005]]

[2]During the 1960s and 70s the International Maize and Wheat Improvement Center (CIMMYT) and the International Rice Research Institute (IRRI) developed high-yielding wheat and rice seeds.

While the importance of TC in agriculture is widely acknowledged in the recent literature [Alston et al., 2009, Huffman and Evenson, 2006, Alene and Coulibaly, 2009, Thirtle et al., 2003], in most agricultural sector models and models of land-use change, TC is still implemented as an exogenous driver [Schneider and Schwab, 2005, Heistermann et al., 2006, Verburg et al., 2009]. In these models, projections primarily depend on a fixed technology path rather than on internal model dynamics. This may lead to serious biases in model results due to an underestimation of the adaptability in the agricultural sector, especially in the longer run. In this chapter I present a first attempt of implementing endogenous technological change in a land-use model, which means that the model can freely decide about the amount of technological change in the future.

The main reason for using an exogenous TC path in most models is that although the relationship between R&D investments in agriculture and technological change is well documented [Alston et al., 2009, Huffman and Evenson, 2006, Alene and Coulibaly, 2009, Thirtle et al., 2003, Federico, 2005, Pardey and Craig, 1989, Alston, 2000], the exact influence of R&D on technological change is still unknown. Several reasons exist for this knowledge gap. First, available time series of R&D investments are still relatively short (less than 30 years) and often incomplete [Pardey and Beintema, 2001]. Second, spillover effects hamper the correct assignment of R&D investments to their impact. Third, success in R&D is hard to predict. High investment may fade away without producing any output, whereas in other instances low investment may create marvelous results. Finally no clear boundary exists between R&D investments in different sectors. In many cases inventions in one sector are based on inventions in other sectors. In a sector analysis of a specific R&D sector, e.g. agricultural R&D, these cross-connections cannot be considered.

In order to deal with these information deficiencies, I have developed a new approach which relates investments in technological change and corresponding yield growth to the τ-factor (Chapter 4). The use of cross-sectional country data in combination with the τ-factor allows to take the land-use intensity of countries as a surrogate for missing time series data. The problems of high uncertainty and unpredictable rates of return associated with investments and the problem of spillovers are partially compensated by using a high aggregation level of only ten world regions[3].

In addition, one can also empirically show that the level of agricultural production costs per area evolves with the τ-factor. Implementing both aspects in a land-use model allows for computing an endogenous pathway of technological change, specifying required TC investments and changes in production costs. The presented experiments in this chapter are performed with the land-use model MAgPIE, which is used for the implementation and validation of the presented method. Future agricultural technology pathways calculated with this approach can be used in other land-use change projections that require external assumptions on technological change.

[3] AFR = Sub-Sahara Africa, CPA = Centrally Planned Asia (incl. China), EUR = Europe (incl. Turkey), FSU = Former Soviet Union, LAM = Latin America, MEA = Middle East and North Africa, NAM = North America, PAO = Pacific OECD (Australia, Japan and New Zealand), PAS = Pacific Asia, SAS = South Asia (incl. India)

5.2 Methodological framework

The internal computation of agricultural technological change is based on the production costs and the effectiveness of R&D investments on yields (investment-yield ratio, IY). The IY ratio evolves with the agricultural land-use intensity, describing the TC investments required per unit of yield growth and area. Accordingly, the production costs are based on the agricultural land-use intensity. For the proposed approach a measure of agricultural land-use intensity is required. For this purpose the τ-factor, which is described in chapter 4, is used. Since the measure is output-related, it captures the full spectrum of yield increasing technology and management options.

5.2.1 Investment-Yield ratio

Based on the τ-factor, it is possible to relate costs to technological change. For this purpose two types of costs have been identified which mainly influence the rate of technological change: first, public and private investments in agricultural R&D, and second, investments in infrastructure (e.g. transport and telecommunication). Data for public and private R&D investments are taken from IFPRI for the year 1981 [Pardey et al., 2006]. Data for infrastructure investments are based on infrastructure costs from the GTAP database, version 7 for the year 2004 [Narayanan and Walmsley, 2008] (discounted from 2004 to 1995)[4]. Unfortunately, the GTAP database does not distinguish between investments in infrastructure and maintenance costs. To get an estimate for annual investments in infrastructure the total GTAP infrastructure costs are corrected with a factor of 0.65 which is the average fraction of investment costs on total infrastructure costs based on OECD [2010]. The remaining 35% of the total infrarstructure costs (the maintenance costs) are treated as additional production costs.

Both investment costs (R&D investments and infrastructure investments) are divided by the average yield growth rate observed in the years 1990-1999 taken from FAO [FAOSTAT, 2009]. The reason for taking the R&D investment data of the year 1981 is the typical time lag between investment in R&D and its impact. The literature offers quite a wide range of various delays and lag-structures proposed for agriculture, ranging from a few years to several decades [Pardey and Craig, 1989, Alston et al., 1998a, Fan et al., 2002, Cox et al., 1997]. I chose a delay of 15 years, which is approximately the average of the delays used in the literature and, according to Alston et al. [1998b] and Alston [2000], the time which is needed to reach the maximum value of gross annual benefits.

As a result, one gets for each region the relationship between investments in agricultural research and the associated yield growth 15 years later. However, the absolute size of investments still depends on the size of a region: the bigger the region, the higher the variation in physical conditions. As a consequence, more research is needed to produce the same average growth rate compared to a smaller region with less variation in physical

[4]Infrastructure investments are composed of investments in transport, water and energy distribution, telecommunication and financial services, all related specifically to the agricultural sector according to GTAP 7 (corresponding GTAP categories names: 'ely', 'gdt', 'wtr', 'cns', 'cmn', 'ofi', 'isr', 'obs', 'ros', and 'osg').

crop conditions. Consequently, I normalized investments relative to the agricultural area of a region. Specific R&D investments per unit of yield growth are computed as the ratio of R&D expenditures per area and the yield growth 15 years later. The same concept is applied for infrastructure investments, except that no time delay is assumed. Both components add up to the investment-yield ratio IY describing the TC investments per area required per human-induced unit yield growth.

To relate this IY ratio to the τ-factor I have calculated the elasticity ϵ_{τ}^{IY}, i.e. the proportional relationship between an increase in the τ-factor and an increase in the IY ratio.

$$\frac{dIY(\tau)}{IY(\tau)} = \epsilon_{\tau}^{IY} \cdot \frac{d\tau}{\tau} \tag{5.1}$$

The elasticity ϵ_{τ}^{IY} is estimated via a regression analysis. Since agricultural R&D data are generally aggregated over all agricultural sectors and spillovers are expected, an aggregated version of the τ-factor covering all crops is used for the regression.

5.2.2 Correlation with production costs

As mentioned above, changes in yield levels are also related to changes in production costs, i.e. costs of all input factors used to produce one unit of output. The initial hypothesis is that production costs per area have a functional relationship with yield level and agricultural land-use intensity. However, since the residuals in the corresponding data are not normally distributed, one cannot use a linear regression analysis to verify the assumption. Instead, I have applied a correlation analysis between (a) yield and costs per area and (b) yield and costs per ton using the Pearson correlation coefficient [Rodgers and Nicewander, 1988] as well as the Kendall rank correlation coefficient [Kendall, 1938]. Two different correlation coefficients are used, to uncover potential, measure-related, biases in the analysis. Whereas the Pearson correlation coefficient measures the magnitude of the linear dependence between two variables, the Kendall rank correlation coefficient measures just any correlation based on a rank test [Kendall and Gibbons, 1990]. Since residuals in the used data set are non-normally distributed, the significance of the Pearson test may be biased, if samples sizes are too small [Kowalski, 1972].

Data for production costs are taken from the GTAP data base, version 7 [Narayanan and Walmsley, 2008], yield data are taken from FAOSTAT [2009]. The data for small producing countries are less accurate and bring much noise into the analysis [Horridge and Laborde, 2008]. Therefore, only the top producing countries for each crop are taken into account so that at least 90% of total crop production is included in the analysis. Another constraint was that at least 1/3 of all available countries (31 countries) are included (an exception is oil palm, which is only produced in 20 countries worldwide).

5.2.3 Model implementation

For the implementation the MAgPIE model is used (Chapter 2.3). Information on model structure, model features and a mathematical description can be found in chapter 2.3. Figure 5.2 shows a schematic overview of the endogenous implementation of technological

change in MAgPIE. Investments in TC lead to a yield increase, which causes the τ-factor to rise. This implies an increase in production costs per area as well as a rise in the IY ratio. Hence, in order to achieve one unit of yield increase in a certain time step, a higher amount of TC investments has to be mobilized than in the previous period.

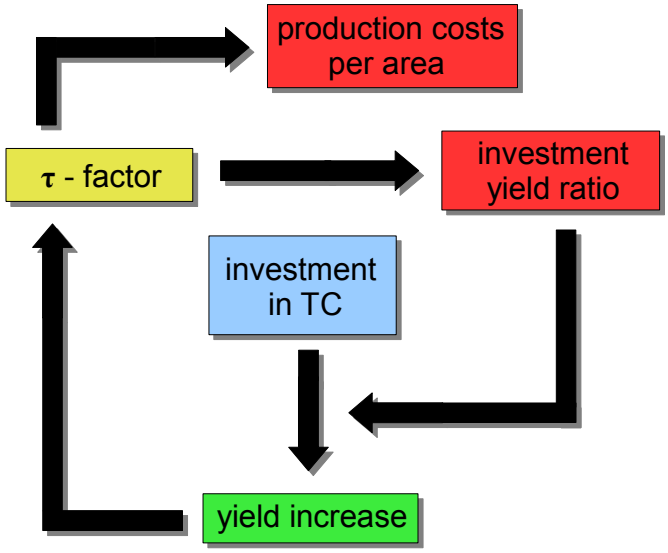

Figure 5.2: Implementation of technological change in MAgPIE (schematic)

In addition, one has to consider some characteristics of the model and the agricultural sector. Typically, for endogenous technology implementations in economic models an intertemporal optimization approach is used due to the need of some kind of planning foresight [Ma and Nakamori, 2009]. In contrast, MAgPIE is a recursive dynamic optimization model which solves each time step separately. To be able to reproduce planning foresight in MAgPIE the annuity approach is used (Section 5.2.4).

Another issue is the implementation of 15 years lag between R&D investment and yield impact. The model decides, based on the expectations 15 years later, how much should be invested. However, since there is no other cross-connection between these time steps, it is possible to shift the investments to the time step when its impact takes place. This means: if the model needs yield growth in the year 2025 due to higher demand expectations, these 2025 model investments must have been made in 2010. However, the costs for R&D in 2010 in the model will be compounded and paid in 2025. This implementation allows for endogenizing technological change in a land-use model without using intertemporal optimization.

5.2.4 Annuity approach

Investments in TC have the characteristic, that they require a high one-time investment, but deliver revenues in more than one successive timestep. For a recursive model, which is accounting for each single timestep on its own, this raises the problem, that the high one-

time investments are out of all proportion to its revenues within a single timestep. This problem can be bypassed by the annuity approach. The annuity approach transfers the lump-sum investments to periodic payments including interest [Kellison, 1991]. Instead of paying the whole investment off in one timestep, the payback is equally split into T payments over the timesteps $t = 1, 2, ..., T$.

The ratio between annual payment C and total investment I_0 is called annuity a (Equation 5.2).

$$C = aI_0 \tag{5.2}$$

For the calculation of annual payments the relation between an investment in timestep t_0 and its value in timestep t_1 becomes important. The value X at timestep t_1 is calculated as the product of the value X at timestep t_0 multiplied with compounding factor $(1+i)^{t_1-t_0}$ with interest rate i and number of intermediate compounding periods $t_1 - t_0$ (Equation 5.3).

$$X_{t_1} = X_{t_0}(1+i)^{t_1-t_0} \tag{5.3}$$

For $t_1 > t_0$ this is the classical compound interest calculation, which is used for example to estimate the future value of a bank account. However, the equation can be used in the same way for calculations backwards in time with $(t_1 < t_0)$.

Having an investment I_0 split in T periodic payments $C_1, C_2, ..., C_T$ combined with equation 5.3 one gets equation 5.4.

$$I_0 = C_1(1+i)^{-1} + C_2(1+i)^{-2} + ... + C_T(1+i)^{-T} \tag{5.4}$$

Assuming equally partitioned periodic payments $C_1 = C_2 = ... = C_T = C$ one gets a geometric progression, which leads to equation 5.5.

$$I_0 = C(1+i)^{-1}\frac{1-(1+i)^{-T}}{1-(1+i)^{-1}} = C\frac{1-(1+i)^{-T}}{i} \tag{5.5}$$

Combining equation 5.5 with equation 5.2 delivers a term describing the annuity (Equation 5.6).

$$a = \frac{i}{1-(1+i)^{-T}} \tag{5.6}$$

In MAgPIE this annuity (Equation 5.6) is multiplied with TC investments (merged to p^{tcc} in Chapter 2.3). Since yield increases gained due to investments in TC typically last forever $T \to \infty$ would deliver a proper accounting for the value of TC. However, to rebuild the observed phenomena of limited foresight and related underinvestment in agricultural TC [Ruttan, 1980, Roseboom, 2002], one can decrease T. For the presented results MAgPIE was running with a limited foresight of $T = 20$ years.

5.2.5 Scenarios

In order to validate the implementation I have compared long-term trends of simulated τ-factor development from 1995 to 2060 with observed data from 1960 to 2005, with a special focus on the overlap in 1995-2005. For the validation historical data from FAO on yield growth is used, which were neither part of the model parametrization nor calibration. Based on this data the changes in τ-factor are calculated backwards starting from 2005.

For the simulation I have applied two scenarios. One scenario which is assuming full protection of intact and frontier forests (IFF) and another scenario without any IFF protection[5]. I used both scenarios because it is hard to judge which scenario is in MAgPIE more close to a "business as usual" case. On one hand the protection of IFF is a manifested objective of many organizations and governments so that investment decisions in R&D more likely be made under the assumption of forest protection. On the other hand deforestation of IFF is happening all over the world and efficient protection mechanism are still lacking. IFF protection is modeled in MAgPIE by excluding the IFF areas from the available land area.

Besides the differences in handling of IFFs both scenarios are based on the same conditions. MAgPIE is driven by external data for population [Center for International Earth Science Information Network (CIESIN) et al., 2000] and gross domestic product (GDP) [World Bank, 2001] (see appendix 2). The food energy demand for the year 1995 is taken from FAOSTAT [2008]. The share of traded goods is kept constant over time and is based on self sufficiency ratios for the year 1995 [FAOSTAT, 2008]. The demand for bioenergy is set in both scenarios to 0.

5.3 Results

5.3.1 Regression and Correlation

The regression result between IY ratio and the τ-factor is the relationship of equation 5.7. Figure 5.3 shows the relationship in a graph for the 10 world regions of MAgPIE.

$$IY(\tau_i) = (3.0 \pm 0.5) \cdot 10^3 \cdot \tau_i^{2.7 \pm 0.9} \tag{5.7}$$

P-values of the t-tests for prefactor a and exponent/elasticity ϵ are $p_a = 0.0005$ (***) and $p_\epsilon = 0.02$ (*). The elasticity between IY ratio and the τ-factor ϵ_τ^{IY} has the value of 2.7 with a standard error of 0.9. As previously explained, changes in τ are proportional to changes in yield, and therefore one can transform this elasticity into an elasticity of yield with respect to accumulated TC investments (I), which is a more common representation

[5]Intact and frontier forests are forests that are mostly undisturbed such as the amazonian rainforest. IFF play an important role in climate change research and ecology because they typically store a huge amount of carbon and contain much biodiversity.

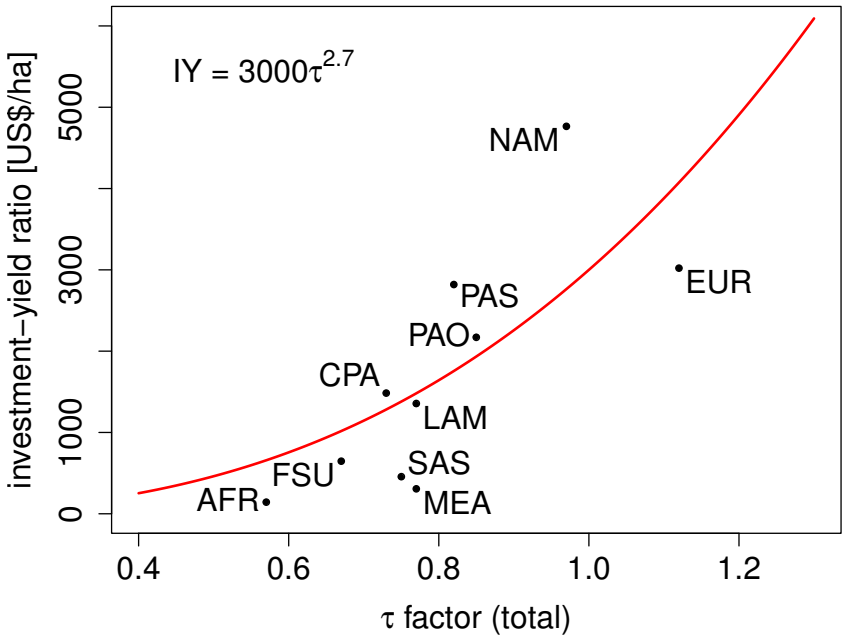

Figure 5.3: investment-yield ratio in relation to τ-factor

(equation 5.8).

$$\epsilon_I^{yld} = \frac{1}{\epsilon_\tau^{IY} + 1} = 0.27 \pm 0.07 \tag{5.8}$$

The result is close to the value of $\epsilon_I^{yld} = 0.296$, as reported in an expert assessment by Nelson et al. [2009].

With regard to the relationship between production costs and yield level, Table 5.2 shows the Pearson correlation coefficients and the Kendall rank correlation coefficients.

All correlations are positive and in most cases at least significant at the 95% level. In the Kendall rank correlation test all crops except tropical cereals, oil palm and sugar cane show significant correlations at the 99.9% significance level. In the pearson correlation tests the results are less significant, but still 10 out of 16 crops show significant correlations at the 95% level. Table 5.3 shows the same information for the relationship between yields and production costs per ton. However, almost none of the tested crop types shows a significant correlation. Comparing the results in both tables suggests the existence of a positive correlation between yields and area-related production costs, but no correlation between yields and output-related production costs. Based on this result production costs per ton have been implemented as a constant input for the model, which leads to a linear increase of production costs per area with yield.

Table 5.4 shows the calculated costs per ton together with the number of countries included in this calculation and the share of total production covered by these countries. These costs per ton are used in MAgPIE for the calculation of production costs (see Chapter 2.3).

crop types		Pearson			Kendall		
		correlation		p-value	correlation		p-value
cereals	temperate	0.81	***	0.000	0.63	***	0.000
	tropical	0.49	*	0.019	0.23		0.140
	maize	0.70	***	0.000	0.61	***	0.000
	rice	0.42	*	0.019	0.57	***	0.000
oilcrops	groundnut	0.17		0.410	0.47	***	0.001
	oil palm	0.07		0.803	0.23		0.228
	rapeseed	0.56	**	0.002	0.55	***	0.000
	soybean	0.08		0.689	0.47	***	0.000
	sunflower	0.68	***	0.000	0.45	***	0.000
sugar	beet	0.65	**	0.002	0.53	***	0.001
	cane	0.37		0.107	0.14		0.422
others	cassava	0.35		0.084	0.47	***	0.001
	potato	0.37	*	0.046	0.58	***	0.000
	pulses	0.75	***	0.000	0.52	***	0.000
	cotton	0.26		0.171	0.49	***	0.000
	others	0.62	***	0.000	0.43	***	0.001

Table 5.2: Correlation between yield and production costs per area (* p≥95%, ** p≥99%, *** p≥99.9%)

crop types		Pearson			Kendall		
		correlation		p-value	correlation		p-value
cereals	temperate	-0.06		0.771	0.15		0.250
	tropical	0.02		0.941	-0.07		0.676
	maize	0.27		0.151	0.25		0.058
	rice	0.28		0.126	0.29	*	0.022
oilcrops	groundnut	-0.10		0.628	0.23		0.118
	oil palm	-0.03		0.912	0.15		0.450
	rapeseed	0.29		0.136	0.26		0.055
	soybean	-0.06		0.753	0.25		0.066
	sunflower	0.12		0.531	0.22		0.103
sugar	beet	0.42		0.068	0.30		0.074
	cane	-0.22		0.352	-0.13		0.461
others	cassava	0.32		0.118	0.25		0.088
	potato	0.22		0.246	0.33	**	0.010
	pulses	0.43	*	0.040	0.38	**	0.010
	cotton	0.00		1.000	0.28	*	0.029
	others	0.42	*	0.025	0.24		0.072

Table 5.3: Correlation between yield and production costs per ton (* p≥95%, ** p≥99%, *** p≥99.9%)

crop types		costs [US$/t]	countries	prod. share
cereals	temperate	130	31	0.95
	tropical	70	31	0.97
	maize	90	31	0.96
	rice	110	31	0.99
oilcrops	groundnut	180	31	1.00
	oil palm	30	20	1.00
	rapeseed	210	31	0.99
	soybean	150	31	1.00
	sunflower	130	31	0.99
sugar	beet	220	31	0.98
	cane	50	31	0.99
others	cassava	350	31	0.99
	potato	1230	31	0.91
	pulses	160	31	0.94
	cotton	620	31	0.99
	others	1130	31	0.92

Table 5.4: Crop-specific, average costs per ton, number of countries used for averaging and the total share of production covered by these countries

5.3.2 Simulation Results

Figure 5.4 shows the development of the τ-factor (2005-2060) compared to past observations of the FAO (1960-2005) in the forest protection scenario. As an example the development path for maize is shown since this is one of the most important crops and is grown in all parts of the world. It is taken as example since all other crop-types in the analysis show similar behavior. Regions like Sub-Saharan Africa (AFR) and North America (NAM) show very strong increases in τ. However, the strongest increase is projected for the Middle East and North Africa region (MEA). This enormous increase is in line with FAO data for this region for the period since the 1980s. Overall three groups can be distinguished: Regions with increasing growth rates (MEA, AFR), constant rates (NAM, LAM, SAS and PAS) and decreasing rates (CPA, EUR, FSU). In this context PAO is a special case with small growth rates in the past but no growth rates in the projections at all.

Figure 5.5 shows the model results of both scenarios compared with FAO observations for the aggregate of all crops. It is important to note that the FAO data used for validation were not used as model input, neither as direct source, nor for calibration purposes. For a direct comparison between observations and model results one can focus on the overlap in 1995-2005. Moreover, the model results can be validated against the general trend in the observed data. For some regions the scenario projections deliver quite similar or even identical results while the projections for other countries strongly depend on the chosen scenario. Especially, the three regions with huge rainforests LAM, AFR, and PAS show high differences in projections. Looking at these three regions also the agree-

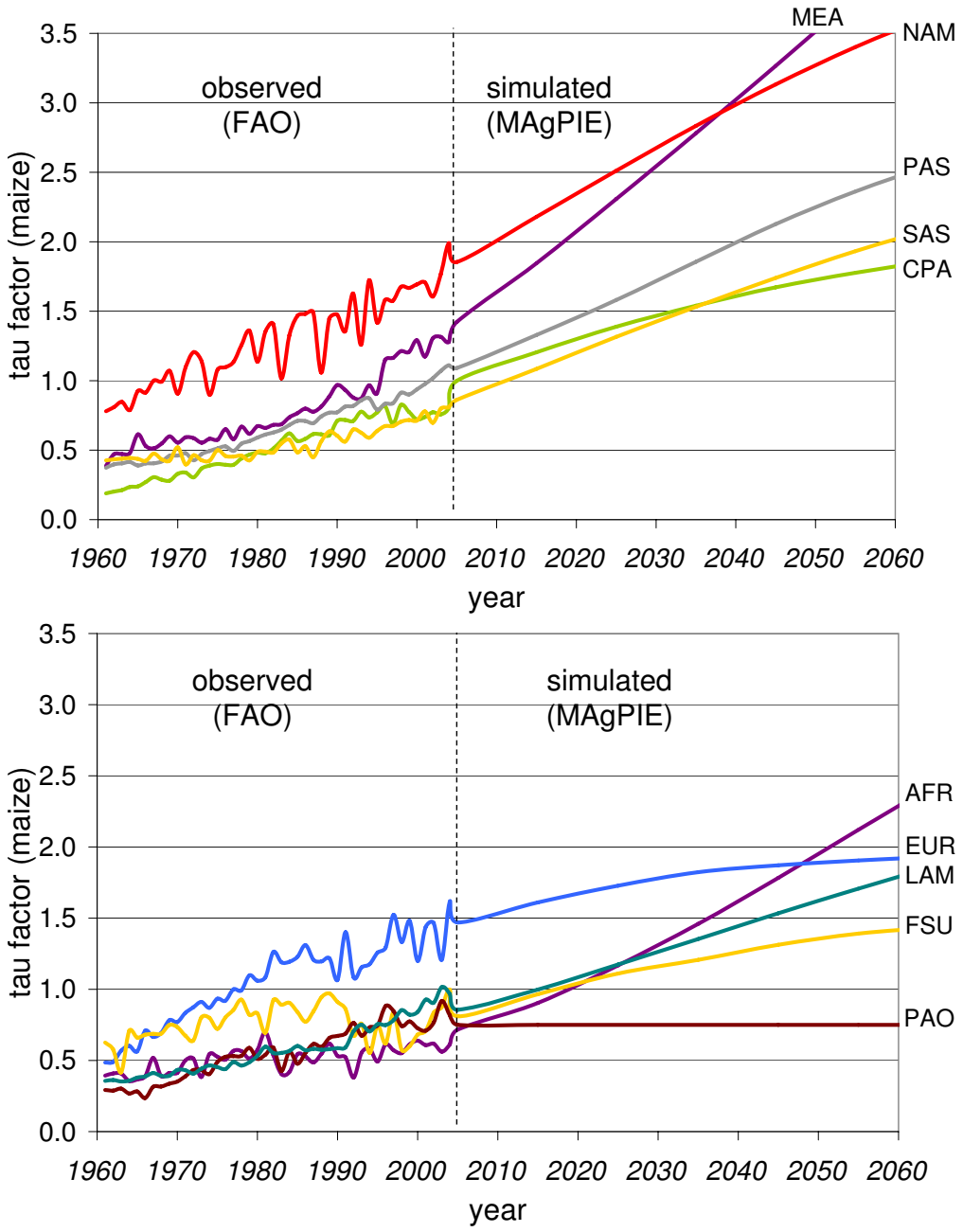

Figure 5.4: Observed and simulated τ-factor for maize in the ten world regions under a forest protection scenario

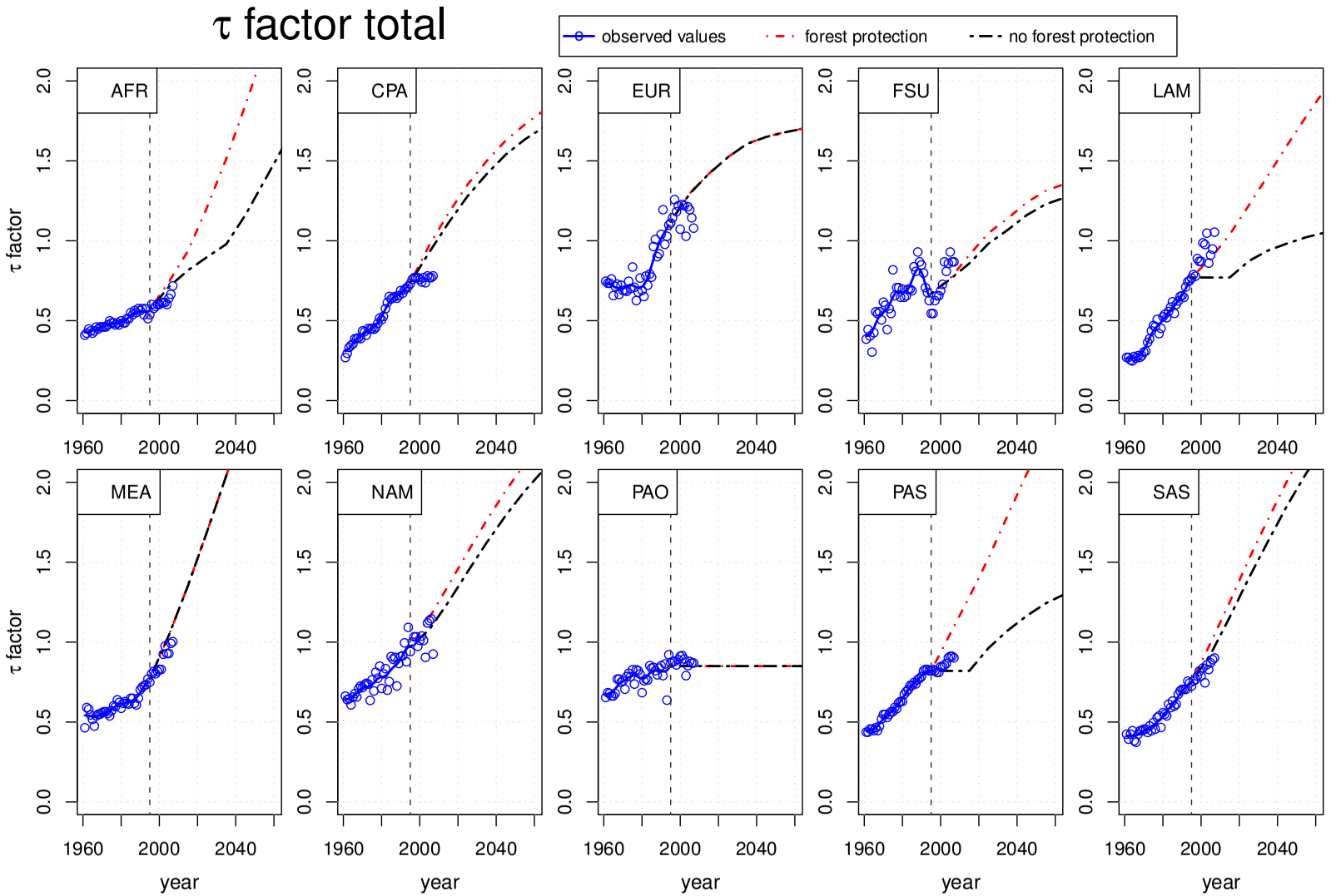

Figure 5.5: Comparison of MAgPIE model projections 1995-2060 in a forest protection scenario (red dotted line) and a scenario without forest protection (black dotted line) with FAO observations 1960-2005 (blue dots) and its running mean (blue line)

ment between observation and validation are quite diverse: In AFR historic growth rates are significantly lower than both projections into the future. However looking at the 10 year overlap the difference is not so big anymore. Whereas the forest protection scenario still projects higher growth rates than observed the scenario without forest protection is quite close to the observations. In contrast, LAM shows the exact opposite behavior. Here the scenario without forest protection significantly underestimates growth rates, while the forest protection scenario reproduces historic trends quite well, but still seems to underestimate the observed growth rates in the overlap. In PAS historic trends fit quite well to the forest protection projection, however in the overlap one can observe some stagnation in the observational data. Surprisingly, the projection without forest protection is showing the same effect, even though more extreme (20 years stagnation instead of only 5 years). This leads to the situation that observed growth rates in PAS lie exactly in the middle of both projections. Looking at the remaining regions the differences between both scenarios are more or less negligible. For EUR, MEA and NAM the general trend as well as the overlap show a good agreement between observation and simulation. In CPA the trend fits well, but in the observed data starting from 1995 one observes a stagnation (similar to the situation in PAS) which is not reproduced by the simulations. The results for FSU is hard to judge because the historic data is strongly affected by fluctuations most likely due to the political transformation after 1990. PAO shows some weak growth in the historic trend but none in the simulation and none in the observed data between 1995-2005. For SAS it seems that both projections slightly overestimate the real trend, even though the difference is only marginal, especially for the case without forest protection. Overall one can say that none of the regions shows dramatic discrepancies between observation and simulation, but for some regions the forest protection scenario shows a better agreement (LAM, PAS) while other regions agree more with the no forest protection scenario (AFR, SAS).

Differences in growth rates between scenarios also directly affect land use patterns. Figure 5.6 and 5.7 show the share of total cropland on total area in 2065 for the forest protection scenario (Figure 5.6) and no forest protection scenario (Figure 5.7). Most differences can be found in the regions LAM, AFR and PAS which were also most sensitive regions in the τ-factor comparison. In these three regions one can see clearly how the rainforests in Brazil, the Democratic Republic of the Congo and Indonesia are cut down. Furthermore, one can find some smaller changes in Canada, Mexico and some shifts in Australia. Due to the absence of relevant IFFs in the rest of the world no other significant changes are taking place.

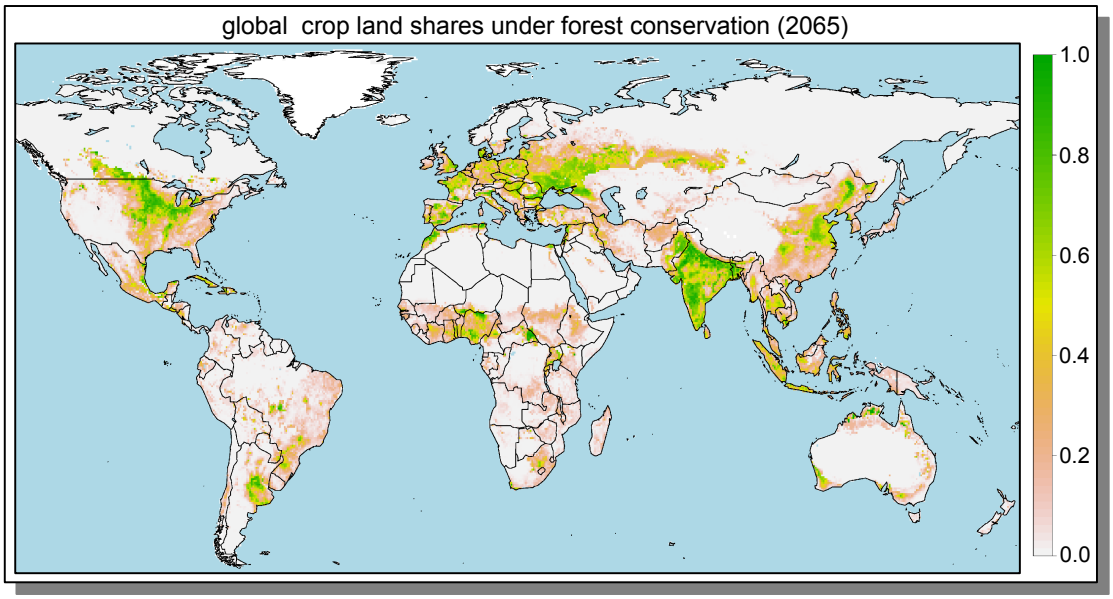

Figure 5.6: Global total cropland shares in a intact and frontier forest protection scenario in 2065

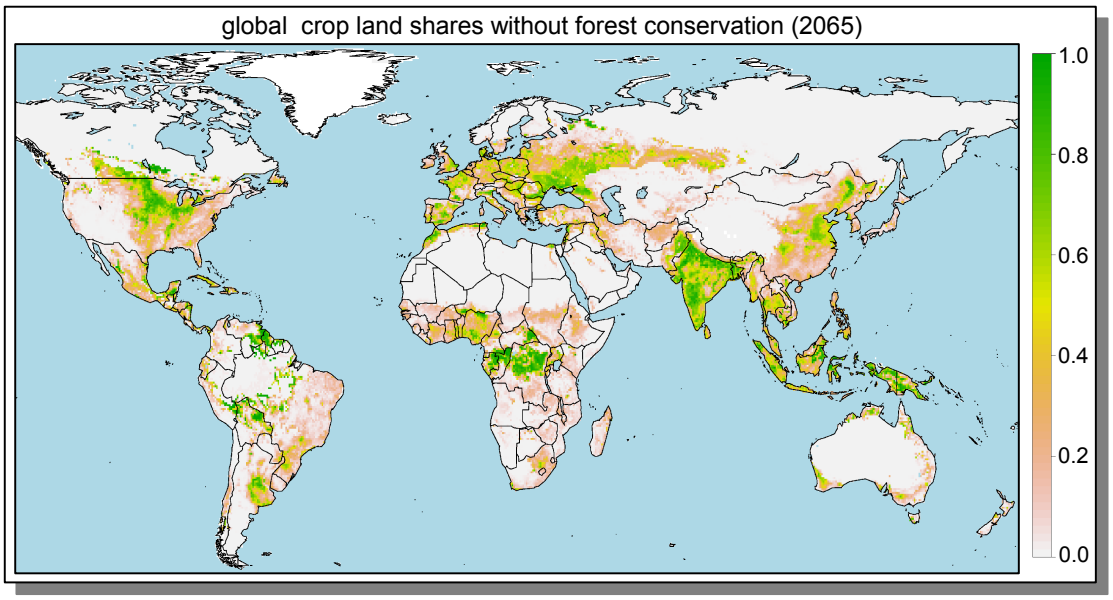

Figure 5.7: Global total cropland shares in absence of any intact and frontier forest protection in 2065

5.4 Discussion

Technological change is a crucial driver for increasing agricultural yields. The presented approach estimates the level and evolution of the investment-yield ratio relative to agricultural land-use intensity. The regression analysis confirms that a higher state of agricultural land-use intensity coincides with a higher IY ratio. Furthermore, the yield elasticity with respect to accumulated TC investments $\epsilon_I^{yld} = 0.27$ is in line with Nelson et al. [2009]. Evenson [1989] showed that R&D spillovers are of major importance in agricultural research if the regions are small. However, increasing the region size, as in the presented implementation, reduces the role of spillover effects significantly. This means that the shown approximation is only valid at coarse scales and becomes invalid applied to finer scales.

Results confirm that yields correlate with production costs per area. Since marginal production costs are constant, every additional production unit costs the same additional amount of money. Consequently, farmers will adopt the new technology since they expect higher yields at constant costs per ton.

The τ projections for maize provide rich insights with regard to future yield trends. The strong increase in Africa indicates what kind of yield growth rates will be required to meet a soaring demand under a forest protection scenario. North America, as the leading region for maize production, continues with high yield growth rates, but could be overtaken by the Middle East and North Africa region (MEA) in terms of growth rates. This region faces unfavorable cropping conditions and at the same time a higher demand increase and a strong political will to reduce food imports. If these conditions prevail in the future, huge investments in technological change would be required. In contrast, Europe continues along its trend over the past two decades when maize yields have not improved much. The Asian regions, starting from a lower yield level and facing a higher demand pressure in the future, have higher growth rates compared to Europe. Lastly, Latin America follows its strong yield growth path since the early 1990s, with high investments in the agricultural sector.

The validation of simulated output with observed data supports the presented model implementation. Especially the long-term trend is reproduced well for most regions, while the observed data in the 1995-2005 overlap often shows some surprising changes in dynamics such as stagnation. A hint for an interpretation of this changes in dynamics can be found in the simulation results of the no forest protection scenario. The projections for LAM as well as for PAS show also a temporary stagnation in growth rates similar to the observed stagnations in CPA and PAS. In the model in this cases additional production is achieved exclusively by land expansion into IFF. However, in both regions the model switches again to yield increases due to technological change. A similar situation could have happened in PAS, which would explain the 5 year stagnation followed up by a continuation in yield growth.

Another interesting aspect is, that AFR is represented best by the scenario without forest protection, LAM by the forest protection scenario and PAS by a mixture of both. This agrees with the political situation in these regions. While LAM is able to trigger investments in R&D on a level which is sufficient to remove the land expansion pressure

based on agricultural demands (there are still other reasons for deforestation), AFR completely fails to do so. PAS seems to have a mixed situation with partial success. The results show that especially in AFR R&D investments have to be increased drastically to achieve food security without cutting down the rainforest in the Congo Basin.

A reason for relatively weak validation results in some regions is that demand and trade are rather inflexible in the current version of MAgPIE and in some cases, like LAM or CPA, this might have strong impacts on future productivity levels. Notwithstanding, the overall validation results indicate the robustness of the approach, since the observed data are not considered as inputs for the analysis and are independent of the model results.

The presented implementation of technological change can be used either in an intertemporal or a recursive dynamic optimization approach. In my case, the latter option is favorable, as it reproduces the observed effect of continuous underinvestment in agricultural R&D [Ruttan, 1980, Roseboom, 2002]. This market failure is caused by the limited foresight of decision makers concerning investments in R&D [Slaughter, 1996]. An intertemporal optimization model, however, would anticipate all the future benefits of R&D investments, which would lead to an optimal R&D investment path in R&D and an overestimation of yield increases, compared with observed trends.

5.5 Conclusion

During the lifetime of Thomas Malthus and before, growth in agricultural output was almost exclusively a result of growth in the use of input factors. This changed by the end of the 19th century and since then agricultural output has been mainly driven by increases in productivity. However, most agricultural sector and land-use models do not cover technological change as an endogenous driver. In order to fill this gap, I have presented a model approach for an endogenous implementation of technological change.

An assumption of the shown implementation, based on numerous studies, is that investments in technological change induce increases in land productivity. The statistical analysis shows that the investment-yield ratio increases in a disproportionate way to the τ-factor and that production costs are linear correlated with the yield level. The results from the model MAgPIE indicate that regions with high demand projections, like Sub-Saharan Africa, or with low potentials for land expansion, like Middle East and South Asia, have to make huge investments in future technological change. While the Middle East region and South Asia show this trend already in observed data, AFR shows this trend only since 1995. Hence, to meet its projected challenges in economic development and meet its growing agricultural demand, it seems indispensable for AFR to increase investments in R&D and infrastructure in order to achieve food security. The endogenous implementation of technological change for reaching food security improves the projection quality of global agricultural models and is a further step towards producing more realistic future scenarios for agriculture.

6 General conclusion

6.1 A brief review

Cross-scale interactions are an important issue in land-use modeling as well as in many other research fields [Wessman, 1992, Cash and Moser, 2000, Harvey, 2000]. Ignoring cross-scale interactions can lead to serious model biases and misleading model dynamics, whereas the inclusion of more detail will strongly affect the computability of a model. Therefore, it is important to find efficient implementations of the relevant cross-scale interactions. Starting with an existing model, as I did, this can be done in two ways: First, improvements of already implemented cross-scale interrelations in terms of accuracy but also computability. Second, inclusion of missing cross-scale dynamics. From this viewpoint I analyzed the agricultural land-use model MAgPIE [Lotze-Campen et al., 2008, 2010, Popp et al., 2010] looking for cross-scale interactions, requiring or allowing improvements most. I found two major starting points, which I have discussed in this thesis: Upscaling of cellular explicit data, so that it becomes utilizable in a global modeling context and technological change as a bidirectional cross-scale interaction between regional and local scale.

In chapter 3 the upscaling issue is addressed and clustering algorithms are proposed as an improvement compared to upscaling using a static grid. The comparison of results produced with a static grid, k-means clustering, hierarchical bottom-up maximal-linkage clustering and a self-developed hierarchical top-down maximal-linkage clustering showed, that clustering is in any case the better choice. Because clustering algorithms are sensitive to the data, that should be upscaled, the same number of final clusters leads to a higher degree of information conservation compared to static grid upscaling. Whereas this general superiority of clustering algorithm was highly significant, the explicit choice of a clustering algorithm was less clear.

Comparing the upscaled data with the original data set k-means showed the best results followed by hierarchical bottom-up clustering. However, for model applications similarity of model outputs produced with upscaled input data to model outputs produced with original data is the relevant factor. The corresponding comparison of model outputs showed that clustering algorithms cannot be ranked easily based on these results and that differences between all upscaling methods diminish compared to the results of the input data comparison. Based on outputs hierarchical bottom-up clustering delivered in most cases the best results. The overall result of this chapter is, that clustering is always a good alternative to static grid upscaling. However, as in many other cluster applications, the explicit choice of a cluster algorithm strongly depends on the explicit problem.

Chapter 4 deals with the topic of agricultural land-use intensity, which is a measure for

the amount of human induced yield increases. The presented development and application of a new measure for agricultural land-use intensity was preparatory work required for the technological change implementation in the following chapter. However, the measure is also quite useful in various contexts different to that. One distinctive feature of the presented approach is, that the measure partly uses model data for calculations. The basic concept is to compare yields, as they are currently observed, with yields, as they would be observed under certain conditions (in the presented case constant agricultural land-use intensities). The hypothetical yields ("reference yields") are estimated by a model (in the presented case LPJmL), whereas the actual yields primarily base on observations.

The presented τ-factor is complementary to most other measures for agricultural land-use intensity. Typically, land-use intensity is measured in an input-based approach, meaning that based on differences in inputs used for agricultural production the agricultural land-use intensity is measured. For instance the concept of cultivation frequencies [Boserup, 2005], which is estimating the land-use intensity based on the use of the input factor land, is following this approach. In contrast, the τ-factor is output-based, meaning that the agricultural outputs (yields) are used for calculations instead of the inputs. Both concepts have advantages and disadvantages. In some cases input data has a better accessibility, in other cases data on outputs is more accessible. Running an input-based approach can deliver useful insights concerning the driver of agricultural land-use intensity: Running an input-based study based on several input factors such as cultivation frequencies or fertilizer use, it is possible to partition the total land-use intensity into parts caused by each of these input factors. On the downside, input-based approaches require detailed knowledge about the relationship of agricultural inputs and outputs. Agricultural land-use intensification is defined as an increase in productivity, therefore a change in inputs is only causing a change in land-use intensity, if also the agricultural output is affected. Though, inaccurate assumptions about input-output relations can lead to serious biases in results. Furthermore input-based approaches are never complete in terms of total agricultural land-use intensity. Only the influence of inputs, that are explicitly taken into account, is measured. In contrast, the τ-factor as an output-based approach does not deliver any information about the different sources of agricultural land-use intensity. The big advantage of the τ-factor is, that it is complete in terms of drivers of agricultural land-use intensity. Any influence on the agricultural land-use intensity is captured, whether caused by a known mechanism or not. Though, output-based measures are not systematically biased towards an underestimation of agricultural land-use intensity. In the calculations a detailed knowledge about input-output relations is substituted by knowledge about the bio-physical processes in agriculture. These characteristics make the τ-factor not in all cases the better choice, but in some. In particular this is the case for an implementation of technological change in a global, agricultural land-use model. For this purpose it is mandatory to have a measure for agricultural land-use intensity, which takes all driving technologies and management improvements into account.

Based on these findings of chapter 4, chapter 5 addressed the bidirectional implementation of technological change (TC) in MAgPIE. Although the implementation is shown

explicitly for the MAgPIE model, the empirical work and its implementation is meant as a general suggestion for agricultural land-use models. The used implementation bases on the idea, that investments decisions for Research & Development (R&D) are taken under consideration of future demand projections. This assumption leads to the situation, that increasing demands can trigger investments in R&D. An important factor for the investment decisions is the investment-yield ratio, which describes the average investment required for a certain yield increase. This ratio is not constant, instead it depends on the yield increase already achieved in the past. To capture the aggregate of past yield increases the τ-factor (Chapter 4) is perfectly equipped. The regression of observed investment-yield ratios with regional τ-factors showed, that increases in agricultural land-use intensity worsen the investment-yield ratio: Higher investments are required to get the same amount of yield increase. Based on this empirically found relationship the TC interaction becomes bidirectional: An investment in R&D leads to yield increases. Concurrently, the agricultural land-use intensity is also increased with the same growth rate, which leads to the feedback of an increasing investment-yield ratio. So, the positive effect of yield increase is accompanied by the negative feedback of increasing investment requirements for further yield increases.

Overall my findings unterpin that apart from detailedness especially the used implementation is of major importance: High-resolution input data is not very useful if most information is lost due to a unfavorable upscaling process. High detailed process implementations do not increase the general model quality if their complexity forces simplifications in other parts of the model such as a reduction in total number of clusters. In the same manner a more detailed reproduction of the technological change implementation on national scale would not deliver better results as it would neglect spillover effects relevant on this scale. Therefore, a good balance between accuracy and abstraction is the most important factor in modelling.

6.2 Future research

6.2.1 Upscaling algorithms

As I showed in chapter 3 clustering algorithms are an improvement compared to upscaling using a static grid. To keep the study concise and to focus on the general implementation issues such as interpolation and choice of data sets for inclusion in the clustering process, I limited myself to the most popular and basic clustering methods k-means and hierarchical clustering. However, the number of available clustering methods is literarily endless and their suitability often strongly depends on the given task [Hartigan, 1985, Jain and Dubes, 1988, pg. 142]. Therefore, it might be interesting to repeat my experiments with an expanded portfolio of clustering techniques.

A candidate with potential for further simulation improvements is bisecting k-means, which is a mixture of the presented methods k-means and hierarchical top-down clustering [Steinbach et al., 2000]. The hierarchical tree is derived by splitting each cluster in two sub-clusters using the k-means algorithm. This combination of both approaches, which are quite distinct, already showed in the study of Steinbach et al. [2000] good re-

sults and might be a superior alternative to the proposed top-down clustering approach. Besides classical clustering techniques it could be also interesting to investigate the performance of algorithms, which are only partly associated with clustering problems, such as self-organizing maps (SOMs) [Kohonen, 1982, 2002]. SOM is a concept for an artificial neural network, which is trained under the use of unsupervised learning. Structural the approach has many similarities to k-means clustering. Both approaches are typically initialized randomly and adapt iteratively the related data. Whereas the basic k-means approach is incorporating in each iteration the full data set, SOMs are only considering one element at once. As SOMs emerged in several studies to be an powerful approach, there might be good chances, that also the presented upscaling problem would benefit of it.

Both candidates are only examples drawn from an endless pool. So, this study could be extended in many directions, which will likely be awarded in many cases with further performance increases.

6.2.2 Uncertainties and Errors

Another aspect related to cross-scale interactions, which was not investigated in my thesis, is the introduction and propagation of errors and uncertainties in a model. The propagation of errors and uncertainties introduced at one scale to another scale is of major relevance for the significance of the various model outputs. Agricultural land-use model inputs are often equipped with some indicators for accuracy, whereas this is typically missing for model outputs. For instance the comparison of different data sets describing the same input parameter can be used as a hint for its accuracy. Another example is the upscaling process explained in chapter 3, for which it is easy to calculate the standard deviation between an element of the original data set and the cluster, which is used as its representation.

As this information is currently not further processed it is hard to judge the robustness of model findings and to distinguish between significant and non-significant outputs. My aim for future research is to tackle this issue. Currently I have two possible approaches in mind: The first solution is to perform a sensitivity analysis of the model, the second one is to estimate output uncertainties based on the propagation of uncertainty concept.

The first solution is to run a sensitivity analysis taking into account the estimated uncertainties of the input parameters. A common approach for a sensitivity analysis is the application of Monte Carlo methods [Kroese et al., 2011]. The idea is to estimate the sensitivity of the model outputs by performing several runs with input parameters, that are randomly altered. In contrast to a systematic alteration of inputs this has the advantage of a reduced amount of required runs and a faster convergence towards the sensitivities of the output variables. However, the required number of runs is still too high to perform a sensitivity analysis over all input parameters for each run of a high-complexity model such as MAgPIE.

Further reduction of sensitivity runs can be achieved by the application of Gaussian quadratures [Stoer and Bulirsch, 2002, pg. 171]. Gaussian quadratures are very effective numerical integration rules. Their special characteristic is the minimal demand of

integration points, which is achieved by the application of orthogonal polynomials. It can drastically reduce the required computation time. In the one-dimensional case n integration points suffice to integrate a polynomial of the order $2n - 1$ exactly.

A sensitivity analysis, as well as the Monte Carlo approach, can also be interpreted as an integration over the input parameter space. Whereas the Monte Carlo method is approximating the integrand by calculating the function values at randomly chosen points, the Gaussian quadrature approach is only requiring the function values at certain, well-defined points (the roots of the used, orthogonal polynomials). This leads to a further, significant performance increase compared to other sensitivity analysis approaches. Nevertheless, the amount of required runs is still to high for an application with each simulation run. Possible solutions for this issue are either a step-wise estimation of uncertainties, with each model run, or a combination of Monte Carlo and Gaussian quadrature concepts. In the first case the idea is to perform with each model run only a few realizations (less than required for an accurate sensitivity analysis). The uncertainties of the model outputs could then be estimated by taking not only the results of the current model run into account, but also of previous model runs. This should lead to results, which are accurate in most cases, at minimal computational costs. However, this approach can deliver misleading results after a model modification. The second approach is currently only an thought, which is not very far developed. One characteristic of a spatial explicit model, as for instance MAgPIE, is, that the model contains several input parameters with similar properties and only differing spatial locations. Under the assumption, that variations in these parameters always deliver similar variations in output parameters a solution would be to aggregate the sensitivity analysis for these elements under use of a Monte Carlo analysis. This would result in a condensed input parameter set, which could then be computed with Gaussian quadratures.

Apart of the sensitivity approach the concept of propagation of uncertainties could also give estimates for the uncertainties of output parameters [Bureau International des Poids et Mesures (BIPM), 1995]. In contrast to the sensitivity analysis no additional model runs would be necessary. Instead uncertainties would be calculated based on a Taylor expansion, which would deliver a simplified version of the used model. For this first or second order approximation of the model dynamics the uncertainties could then be calculated analytically. A problem of this concept is, that it would require a lot of additional calculations the modelers have to perform manually and the results would be more error-prone compared to a sensitivity analysis.

All approaches have in common, that they can only provide indicators for uncertainties of the different outputs. This has several reasons: First, provided uncertainty values of inputs are typically only an estimate. Second, not all uncertainties can be taken into account. False model implementations or missing links are also an significant source for model biases. Third, the proposed analysis methods are also only estimates. Nevertheless, having in mind, that the calculated values are just indicators for uncertainties and biases, they can be quite useful and would be a significant advancement to the current situation.

Supplementary data - Clustering

1 Implementation hierarchical top-down clustering

```
"""#### Implementation of hierarchical top-down clustering
      as used in MAgPIE"""

version = 3.7

from numpy import array, copy, lexsort, argsort
from cPickle import dump
from struct import Struct
from Pycluster import distancematrix
from itertools import islice, ifilter
import heapq, os

class DistFile:
    '''DistFile Class: Generates a DistanceFileObject containing elements
    of the structure (distance,left,right) - Object can be created using
    either a distancematrix or a list of sorted distance elements'''

    p = Struct('fii') #"struct" object containing encoding (float,int,int)
    elemsize = 12     #element size in bytes
    count = 0         #general object counter
    filename = "distfile%i.tmp"  #file name under which data is stored
    #maximal number of elements sorted directly
    #in the case of files with more elements sorting
    #procedure is splitted into parts
    maxsortlength = 10**6
    distscale = 10**11  #distance scaling factor

    def __init__(self, input=None, offset=0):
        DistFile.count += 1
        self.filename = DistFile.filename % DistFile.count
        self.length = 0
        if(input!=None):
            if(type(input)==str):
                self.load_file(input)
            else:
                try:
                    is_dmatrix = (len(input[0])==0)
                except:
                    is_dmatrix = False
                if is_dmatrix:
                    self.dmatrix_write(input, offset)
                    self.sort()
                else:
```

```
                self.dist_write(input)

    def __del__(self):
        try:
            os.remove(self.filename)
        except:
            pass

    def __getitem__(self, i):
        if isinstance(i, slice):
            return list(islice(self.read(), i.start, i.stop, i.step))
        else:
            return list(islice(self.read(), i, i+1))

    def __iter__(self):
        return self.read()

    def __len__(self):
        return self.length

    def dmatrix_write(self, dmatrix, offset=0):
        with open(self.filename, 'wb') as dfile:
            for i in range(0, len(dmatrix)-1):
                for j in range(i+1, len(dmatrix)):
                    dfile.write(self.p.pack(-dmatrix[j][i]*
                                DistFile.distscale, i+offset, j+offset))
        self.length = (len(dmatrix)**2 - len(dmatrix))/2

    def dist_write(self, dist):
        self.length = 0
        with open(self.filename, 'wb') as dfile:
            for elem in dist:
                dfile.write(self.p.pack(elem[0], elem[1], elem[2]))
                self.length += 1

    def merge_write(self, *distfilelist):
        self.length = 0
        with open(self.filename, 'wb') as dfile:
            for elem in heapq.merge(*distfilelist):
                dfile.write(self.p.pack(elem[0], elem[1], elem[2]))
                self.length += 1

    def load_file(self,filename):
        self.filename = filename
        self.length = int(os.path.getsize(self.filename))/DistFile.elemsize

    def append(self, elem):
        with open(self.filename, 'ab') as dfile:
            dfile.write(self.p.pack(elem[0], elem[1], elem[2]))
            self.length += 1

    def extend(self, dist):
        with open(self.filename, 'ab') as dfile:
            for elem in dist:
```

```
                dfile.write(self.p.pack(elem[0], elem[1], elem[2]))
        self.length += len(dist)


    def read(self):
        try:
            with open(self.filename, 'rb') as dfile:
                try:
                    while True:
                        yield self.p.unpack(dfile.read(12))
                except:
                    pass
        except:
            pass

    def filter_read(self,key):
        for elem in ifilter(key,self.read()):
            yield elem

    def sort(self):
        if self.length <= DistFile.maxsortlength:
            self.dist_write(sorted(self))
        else:
            pos = 0
            tmpdists = []
            while True:
                print "Create temporary distfile for sorting!"
                tmpdists.append(DistFile())
                tmpdists[-1].dist_write(sorted(self[pos:pos +
                                                    DistFile.maxsortlength]))
                pos += DistFile.maxsortlength
                if pos >= self.length:
                    break
            self.merge_write(*tmpdists)
            for elem in tmpdists:
                del(elem)

def splitcluster(distfile, cells, counter=0):
    '''function used to split one cluster into two sub-clusters
    distfile is a DistanceFile containing all distances between
    cells of the cluster, cells contains the cells that are part
    of the cluster'''

    len_cells = len(cells)

    print "start splitcluster with",len_cells,"cells"

    #sort array based on distances
    if len(distfile) == 0: #single cell -> no splitting
        return [counter, None]
    if len(distfile) == 1: #two cells -> trivial case
        return [counter-1, list(distfile[0])]

    #open two list starting with first cells of darray (longest distance)
```

```
#structure key : (pairingblock : left/right) (left/right = False/True)
#each pairingblock describes which cells will definitively be in
#different clusters at the end only one pairingblock will remain,
#containing the cells of both clusters
ddict = {}

pairingcount = 0
recombinecount = 0
loopcounter = 0
last_ddict_len = 0
cells = set(cells)

for elem in distfile:

    if (elem[1] not in ddict) and (elem[2] not in ddict):
        #create new pairing, because both entries do not exist so far
        print "create new pair ", pairingcount
        ddict[elem[1]] = [pairingcount, False]
        ddict[elem[2]] = [pairingcount, True]
        pairingcount += 1

    elif (elem[1] in ddict) and (elem[2] not in ddict):
        #First entry is already in dictionary, but second entry is not
        #-> combine based on first entry
        ddict[elem[2]] = [ddict[elem[1]][0], not ddict[elem[1]][1]]

    elif (elem[1] not in ddict) and (elem[2] in ddict):
        #Second entry is already in dictionary, but first entry is not
        # -> combine based on second entry
        ddict[elem[1]] = [ddict[elem[2]][0], not ddict[elem[2]][1]]

    #both entries are already in the dictionary,
    #but different pairingblocks

    #combine blocks
    elif (ddict[elem[1]][0] != ddict[elem[2]][0]):
        #sides of one pairingblock have to be switched if both
        #entries are on the same side
        switchside = (ddict[elem[1]][1] == ddict[elem[2]][1])
        remainingblock = min(ddict[elem[1]][0], ddict[elem[2]][0])
        removingblock = max(ddict[elem[1]][0], ddict[elem[2]][0])
        recombinecount += 1
        for key, cell in ddict.iteritems():
            if cell[0] == removingblock:
                ddict[key][0] = remainingblock
                if(switchside):
                    ddict[key][1] = (not ddict[key][1])


    #exit iteration, when every cell is mapped to one sub-cluster
    if (len(ddict)==len_cells) and (pairingcount==recombinecount+1):
        break

    loopcounter += 1
```

```
        if(len(ddict)%100==0 and len(ddict) > last_ddict_len):
            last_ddict_len = len(ddict)
            print ("Loop",loopcounter, "len(ddict)=",len(ddict),
                   "len(cells)=",len_cells)

    print ("all cells mapped to a cluster", len_cells,
           "len(ddict)=",len(ddict))

    #save first element as topelem
    topelem = distfile.read().next()
    print "got first element", topelem

    left_cells = set(filter(lambda x: ddict[x][1]==False,ddict.keys()))
    right_cells = set(filter(lambda x: ddict[x][1]==True,ddict.keys()))

    print "cells filtered"

    #left branch
    left_dist = ifilter(lambda x: (x[1] in left_cells)
                        and (x[2] in left_cells),distfile)
    print "left filter ready"
    distfile_left = DistFile(left_dist)
    print "distfile left ready"

    [counter, temp] = splitcluster(distfile_left, left_cells, counter)
    del(distfile_left)
    if(temp == None):
        leftcluster = topelem[1]
        treearray = list()
    else:
        leftcluster = counter
        treearray = temp

    #right branch
    right_dist = ifilter(lambda x: (x[1] in right_cells) and
                         (x[2] in right_cells),distfile)
    print "right filter ready"
    distfile_right = DistFile(right_dist)
    print "distfile right ready"
    [counter, temp] = splitcluster(distfile_right, right_cells, counter)
    del(distfile_right)
    if(temp == None):
        rightcluster = topelem[2]
    else:
        rightcluster = counter
        treearray.extend(temp)

    treearray.append((topelem[0], leftcluster, rightcluster))

    return [counter-1, treearray]

def write_tree(tree_array,filename):
    fileHandle = open(filename, "w")
```

```
    dump(tree_array,fileHandle)
    fileHandle.close()

def hierarchical_topdown(data, cpr=None,filename="cluster_c.tree"):
    if cpr == None:
        cpr = [data.shape[0]]
    cells_used = 0
    linecount = 0
    treearray = list()
    start = 0
    combineregions = False

    for ncells in cpr:
        stop = start + ncells
        print "General DistFile for a region!"
        dist = DistFile(distancematrix(data[start:stop]),offset=cells_used)
        #start splitcluster to get an unordered top-down tree
        counter, treearraypart = splitcluster(dist, range(start,stop),
                                              counter=-linecount)
        start = stop
        treearray.extend(treearraypart)

        f = open("treearray%i.dat"%linecount,"w")
        dump(treearray,f)
        f.close()
        print treearray

        prev_linecount = linecount
        linecount += ncells-1
        if combineregions:
            #add connectors between regions, set distance to -999
            treearray.append(array([999, -prev_linecount, -linecount]))
            linecount += 1
        cells_used += ncells
        combineregions = True
    treearray = array(treearray)
    treearray[:, 0] = -treearray[:, 0]
    #set distances between regions higher than distances within regions
    maxdist = max(treearray[:, 0])
    treearray[treearray[:, 0] == -999, 0] = maxdist + 100
    #calculate sortorder based on distances
    order = lexsort((-treearray[:,2],-treearray[:,1],treearray[:, 0]))
    treearray = treearray[order, :]      #sort array
    #remove fake distances
    treearray[treearray[:, 0] > maxdist, 0] = None
    treearray_old = copy(treearray)      #create a copy
    #relabel clusters based on new order (clusters are named based on
    #the iteration they are created
    for i in range(-1, -(treearray.shape[0]+1), -1):
        treearray[treearray_old[:, 1] == i, 1] = -(argsort(order)[-i-1]+1)
        treearray[treearray_old[:, 2] == i, 2] = -(argsort(order)[-i-1]+1)

    write_tree(treearray,filename)
```

2 Quality results for component runs

Table 1: Quality of input and output data of selected upscalings using hierarchical bottom-up clustering with 400 (h400) and 1438 (h1438) cells and varying numbers and types of data sets used for upscaling: "all data" = All available data sets, "input/output best/worst": A combination of that 10 data sets showing the best/worst performance in input/output quality measures, when only using that single data set for upscaling. "yield_rf": rainfed yields. "yield_ir": irrigated yields. "airrig": annual water demand for irrigation.

		input				**output**			
		d_1		d_2		d_1		d_2	
		h400	h1438	h400	h1438	h400	h1438	h400	h1438
all data		0.24	0.11	0.24	0.16	0.70	0.63	0.79	0.76
input	**best**	0.38	0.28	0.32	0.23	0.80	0.73	0.84	0.81
	worst	0.63	0.53	0.43	0.33	0.74	0.69	0.81	0.79
output	**best**	0.49	0.33	0.40	0.29	0.74	0.59	0.81	0.75
	worst	0.31	0.19	0.28	0.21	0.77	0.71	0.83	0.81
yield_rf	**wheat**	0.69	0.64	0.74	0.78	0.81	0.72	0.90	0.87
	maize	0.86	0.85	0.67	0.69	0.78	0.70	0.88	0.84
	millet	0.66	0.68	0.65	0.62	0.79	0.68	0.90	0.84
	rice	0.70	0.62	0.70	0.65	0.79	0.67	0.88	0.85
	soybean	0.75	0.71	0.63	0.66	0.77	0.71	0.90	0.87
	rapeseed	0.70	0.66	0.76	0.72	0.81	0.73	0.91	0.90
	groundnut	0.71	0.64	0.73	0.66	0.79	0.71	0.86	0.86
	sunflower	0.81	0.68	0.65	0.62	0.80	0.71	0.90	0.88
	oil palm	0.70	0.64	0.72	0.67	0.78	0.68	0.86	0.84
	pulses	0.75	0.60	0.68	0.63	0.79	0.71	0.90	0.87
	potato	0.63	0.53	0.60	0.62	0.76	0.69	0.90	0.88
	cassava	0.73	0.66	0.75	0.69	0.78	0.70	0.87	0.86
	sugar cane	0.71	0.65	0.73	0.67	0.77	0.72	0.85	0.85
	sugar beet	0.63	0.54	0.61	0.62	0.75	0.71	0.89	0.89
	others	0.86	0.77	0.90	0.73	0.71	0.71	0.87	0.92
	cotton	0.75	0.62	0.67	0.67	0.79	0.70	0.89	0.87

Table 1: (continued) Quality of input and output data of selected upscalings using hierarchical bottom-up clustering with 400 (h400) and 1438 (h1438) cells and varying numbers and types of data sets used for upscaling: "all data" = All available data sets, "input/output best/worst": A combination of that 10 data sets showing the best/worst performance in input/output quality measures, when only using that single data set for upscaling. "yield_rf": rainfed yields. "yield_ir": irrigated yields. "airrig": annual water demand for irrigation.

		input				**output**			
		d_1		d_2		d_1		d_2	
		h400	h1438	h400	h1438	h400	h1438	h400	h1438
	fodder	0.81	0.75	0.76	0.82	0.86	0.79	0.94	0.90
	pasture	0.81	0.75	0.76	0.82	0.86	0.79	0.94	0.90
	bioen. grasses	0.61	0.55	0.65	0.69	0.78	0.75	0.89	0.88
	bioen. trees	0.75	0.73	0.73	0.73	0.83	0.82	0.91	0.92
yield_ir	**wheat**	0.64	0.59	0.62	0.65	0.79	0.69	0.89	0.86
	maize	0.60	0.56	0.59	0.62	0.82	0.76	0.90	0.89
	millet	0.63	0.54	0.59	0.55	0.79	0.68	0.90	0.84
	rice	0.68	0.64	0.70	0.66	0.79	0.74	0.90	0.89
	soybean	0.61	0.56	0.59	0.58	0.81	0.71	0.90	0.86
	rapeseed	0.64	0.61	0.63	0.62	0.77	0.71	0.89	0.88
	groundnut	0.69	0.63	0.70	0.63	0.80	0.70	0.90	0.87
	sunflower	0.62	0.58	0.59	0.59	0.80	0.76	0.91	0.89
	oil palm	0.68	0.61	0.70	0.60	0.79	0.71	0.90	0.87
	pulses	0.60	0.55	0.56	0.57	0.80	0.72	0.89	0.89
	potato	0.59	0.53	0.54	0.57	0.80	0.80	0.89	0.91
	cassava	0.73	0.64	0.76	0.63	0.78	0.72	0.88	0.87
	sugar cane	0.69	0.62	0.71	0.63	0.80	0.73	0.92	0.88
	sugar beet	0.58	0.53	0.53	0.58	0.81	0.79	0.89	0.89
	others	0.86	0.77	0.90	0.73	0.71	0.71	0.87	0.92
	cotton	0.60	0.54	0.56	0.57	0.79	0.72	0.90	0.89
	fodder	0.66	0.57	0.56	0.57	0.82	0.79	0.90	0.90
	pasture	0.66	0.57	0.56	0.57	0.82	0.79	0.90	0.90
	bioen. grasses	0.59	0.57	0.61	0.72	0.80	0.78	0.91	0.91
	bioen. trees	0.64	0.62	0.64	0.69	0.79	0.78	0.89	0.90
airrig	**wheat**	0.59	0.56	0.56	0.61	0.79	0.80	0.91	0.92
	maize	0.55	0.52	0.49	0.53	0.83	0.82	0.90	0.92

Table 1: (continued) Quality of input and output data of selected upscalings using hierarchical bottom-up clustering with 400 (h400) and 1438 (h1438) cells and varying numbers and types of data sets used for upscaling: "all data" = All available data sets, "input/output best/worst": A combination of that 10 data sets showing the best/worst performance in input/output quality measures, when only using that single data set for upscaling. "yield_rf": rainfed yields. "yield_ir": irrigated yields. "airrig": annual water demand for irrigation.

	input				**output**			
	d_1		d_2		d_1		d_2	
	h400	h1438	h400	h1438	h400	h1438	h400	h1438
millet	0.55	0.51	0.48	0.52	0.81	0.83	0.90	0.92
rice	0.55	0.51	0.48	0.52	0.80	0.80	0.91	0.91
soybean	0.55	0.54	0.50	0.56	0.80	0.79	0.90	0.89
rapeseed	0.59	0.57	0.57	0.64	0.83	0.79	0.91	0.92
groundnut	0.55	0.50	0.48	0.52	0.80	0.81	0.90	0.92
sunflower	0.56	0.53	0.52	0.55	0.81	0.80	0.90	0.90
oil palm	0.55	0.50	0.48	0.52	0.80	0.81	0.90	0.92
pulses	0.59	0.56	0.56	0.62	0.82	0.81	0.91	0.92
potato	0.59	0.55	0.52	0.58	0.82	0.82	0.91	0.92
cassava	0.56	0.51	0.49	0.52	0.82	0.79	0.91	0.91
sugar cane	0.55	0.50	0.48	0.52	0.80	0.81	0.90	0.92
sugar beet	0.59	0.55	0.52	0.58	0.82	0.82	0.91	0.92
others	0.86	0.77	0.90	0.73	0.71	0.71	0.87	0.92
cotton	0.59	0.56	0.56	0.62	0.82	0.81	0.91	0.92
fodder	0.58	0.54	0.52	0.57	0.80	0.81	0.90	0.91
pasture	0.58	0.54	0.52	0.57	0.80	0.81	0.90	0.91
bioen. grasses	0.60	0.58	0.58	0.64	0.80	0.76	0.89	0.88
bioen. trees	0.61	0.59	0.62	0.75	0.78	0.76	0.89	0.90

Country-to-region mapping

AFR **Sub-Saharan-Africa**	**CPA** **Centr. Planned Asia**	**EUR** **Europe**	**FSU** **Former Soviet Union**	**LAM** **Latin America**
Angola	Cambodia	Albania	Azerbaijan, Republic of	Argentina
Benin	China	Austria	Belarus	Belize
Botswana	Laos	Belgium-Luxembourg	Georgia	Bolivia
Burkina Faso	Mongolia	Bosnia and Herzegovina	Kazakhstan	Brazil
Burundi	Viet Nam	Bulgaria	Kyrgyzstan	Chile
Cameroon		Croatia	Moldova, Republic of	Colombia
Central African Republic		Czech Republic	Russian Federation	Costa Rica
Chad		Denmark	Tajikistan	Cuba
Congo, Dem Republic of		Estonia	Turkmenistan	Dominican Rep.
Congo, Republic of		Finland	Ukraine	Ecuador
Cote d'Ivoire		France	Uzbekistan	El Salvador
Djibouti		Germany		French Guiana
Equatorial Guinea		Greece		Guatemala
Eritrea		Hungary		Guyana
Ethiopia		Iceland		Haiti
Gabon		Ireland		Honduras
Ghana		Italy		Mexico
Guinea		Latvia		Nicaragua
Guinea-Bissau		Lithuania		Panama
Kenya		Macedonia		Paraguay
Lesotho		Netherlands		Peru
Liberia		Norway		Suriname
Madagascar		Poland		Uruguay
Malawi		Portugal		Venezuela
Mali		Romania		
Mauritania		Slovakia		
Mozambique		Slovenia		
Namibia		Spain		
Niger		Sweden		
Nigeria		Switzerland		
Rwanda		Turkey		
Senegal		United Kingdom		
Sierra Leone		Yugoslavia, Fed Rep of		
Somalia				
South Africa				
Sudan				
Swaziland				
Tanzania, United Rep. of				
Togo				
Uganda				
Western Sahara				
Zambia				
Zimbabwe				
MEA **Middle East/North Afr.**	**NAM** **North America**	**PAO** **Pacific OECD**	**PAS** **Pacific Asia**	**SAS** **South Asia**
Algeria	Canada	Australia	Indonesia	Afghanistan
Egypt	USA	Japan	Korea, Dem People's Rep	Bangladesh
Iran, Islamic Rep of		New Zealand	Korea, Rep of	Bhutan
Iraq			Malaysia	India
Israel			Papua New Guinea	Myanmar
Jordan			Philippines	Nepal
Kuwait			Solomon Islands	Pakistan
Libyan Arab Jamahiriya			Thailand	Sri Lanka
Morocco				
Oman				
Saudi Arabia				
Syrian Arab Rep				
Tunisia				
United Arab Emirates				
Yemen				

Table 2: Country-to-region mapping

Population and GDP assumptions

year	AFR	CPA	EUR	FSU	LAM	MEA	NAM	PAO	PAS	SAS
1995	553	1281	554	276	452	278	292	134	383	1270
2005	743	1480	589	293	550	357	332	146	462	1572
2015	926	1582	586	295	623	423	355	148	517	1797
2025	1125	1651	575	295	687	486	375	147	565	1998
2035	1313	1673	559	285	739	541	391	146	614	2149
2045	1481	1677	532	275	780	590	400	144	652	2265
2055	1629	1659	505	262	810	633	404	140	674	2347
2065	1753	1632	480	246	830	671	403	132	684	2398
2075	1845	1610	458	232	844	701	402	122	690	2423
2085	1914	1599	449	224	855	728	401	112	685	2440
2095	1953	1590	440	216	861	752	400	100	676	2452

Table 3: Population in million people from 1995 to 2095 aggregated to ten world regions [Center for International Earth Science Information Network (CIESIN) et al., 2000]

year	AFR	CPA	EUR	FSU	LAM	MEA	NAM	PAO	PAS	SAS
1995	1513	3299	16128	3521	6527	4940	26765	21469	3649	1461
2005	1627	5855	20124	4081	7840	5855	33920	24240	4614	2139
2015	1826	8907	25189	6094	9769	7352	39349	28672	6692	3180
2025	2080	12311	30654	8496	11853	9215	44489	34841	9324	4406
2035	2447	16270	36115	11143	14131	11408	49842	41224	12371	5805
2045	3221	20512	41080	15264	17144	14142	55597	45297	16211	7769
2055	4242	24720	45851	20235	20808	17346	61383	49037	20322	9827
2065	5430	28579	50672	25698	24989	21002	67106	52935	24569	11923
2075	6823	32461	55419	31465	29688	25101	72804	56813	29050	14083
2085	8425	36296	60119	37247	35057	29530	78330	60693	33879	16244
2095	10299	40273	65026	42963	41189	34405	84071	64730	39281	18557

Table 4: GDP per capita (US$ per number of people in purchasing power parities (PPP)) [World Bank, 2001]

Bibliography

D. M. Adams, R. J. Alig, J. M. Callaway, B. A. McCarl, and S. M Winnett. *The forest and agricultural sector optimization model (FASOM): model structure and policy applications.* U.S. Dept. of Agriculture, Forest Service, Pacific Northwest Research Station, 1996.

A. D. Alene and O. Coulibaly. The impact of agricultural research on productivity and poverty in sub-Saharan africa. *Food Policy*, 34(2):198–209, 2009.

J. M. Alston. *A meta-analysis of rates of return to agricultural R&D: Ex pede Herculem?* Int Food Policy Res Inst IFPRI, 2000.

J. M. Alston, B. J. Craig, and P. G. Pardey. *Dynamics in the Creation and Depreciation of Knowledge and the Returns to Research.* Environment and Production Technology Division, International Food Policy Research Institute, 1998a.

J. M. Alston, G. W. Norton, and P. G. Pardey. *Science under scarcity. Principles and practice for agricultural research evaluation and priority setting.* CAB International, 1998b.

J. M. Alston, J. M. Beddow, and P. G. Pardey. Agricultural research, productivity, and food prices in the long run. *Science*, 325(5945):1209–1210, 2009.

S. S. Bakken, A. Aulbach, E. Schmid, J. Winstead, L. T. Wilson, R. Lerdorf, and Z. Suraski. *PHP manual.* PHP Documentation Group, 2000.

M. J. Bellenger and A. T. Herlihy. An economic approach to environmental indices. *Ecological Economics*, 68(8-9):2216–2223, 2009.

G. Blöschl and M. Sivapalan. Scale issues in hydrological modelling: a review. *Hydrological processes*, 9(3-4):251–290, 1995.

A. Bondeau, P.C. Smith, S. O. N. Zaehle, S. Schaphoff, W. Lucht, W. Cramer, D. Gerten, H. Lotze-Campen, C. Müller, and M. Reichstein. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biology*, 13(3): 679–706, 2007.

E. Boserup. *The conditions of agricultural growth: the economics of agrarian change under population pressure.* Aldine De Gruyter, 2005.

A. Brook, D. Kendrick, and A. Meeraus. GAMS, a user's guide. *ACM SIGNUM Newsletter*, 23(3-4):10–11, 1988.

H. C. Brookfield. Notes on the theory of land management. *PLEC News and Views*, 1: 28–32, 1993.

J. Bruinsma. *World agriculture: towards 2015/2030: an FAO perspective*. Earthscan/-James & James, 2003.

Bureau International des Poids et Mesures (BIPM). *Guide to the expression of uncertainty in measurement*, 1995. URL `http://www.bipm.org/en/publications/guides/gum.html`.

D. W. Cash and S. C. Moser. Linking global and local scales: designing dynamic assessment and management processes. *Global Environmental Change*, 10(2):109–120, 2000.

M. Cassel-Gintz and G. Petschel-Held. GIS-based assessment of the threat to world forests by patterns of non-sustainable civilisation nature interaction. *Journal of Environmental Management*, 59(4):279–298, 2000.

Center for International Earth Science Information Network (CIESIN), International Policy Research Institute (IFPRI), and World Research Institute (WRI). Gridded population of the world, Version 2, 2000. URL `http://sedac.ciesin.columbia.edu/plue/gpw`.

T. J. Coelli and D. S. P. Rao. Total factor productivity growth in agriculture: a malmquist index analysis of 93 countries, 1980-2000. *Agricultural Economics*, 32(s1): 115–134, 2005.

T. Cox, J. Mullen, and W. Hu. Nonparametric measures of the impact of public research expenditures on australian broadacre agriculture. *The Australian Journal of Agricultural and Resource Economics*, 41(3):333–360, 1997.

M. J. L. de Hoon, S. Imoto, J. Nolan, and S. Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

J. P. Dietrich. Phase space reconstruction using the frequency domain : a generalization of actual methods. Master's thesis, University of Potsdam, 2008. URL `http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-50738`.

J. P. Dietrich, A. Popp, and H. Lotze-Campen. Reducing the loss of information and gaining accuracy with clustering methods in a global land-use model. *Environmental Modelling & Software (submitted)*, 2011a.

J. P. Dietrich, C. Schmitz, H. Lotze-Campen, A. Popp, and C. Müller. Forecasting technological change in agriculture - an endogenous implementation in a global land use model. *Technological Forecasting & Social Change (submitted)*, 2011b.

J. P. Dietrich, C. Schmitz, C. Müller, M. Fader, H. Lotze-Campen, and A. Popp. Measuring agricultural land-use intensity - a global analysis using a model-assisted approach. *Ecological Modelling (submitted)*, 2011c.

T. Dirnböck, P. Bezák, S. Dullinger, H. Haberl, H. Lotze-Campen, M. Mirtl, J. Peterseil, S. Redpath, S. J. Singh, J. Travis, and S. M. J. Wijdeven. Scaling issues in long-term socio-ecological biodiversity research: A review of european cases. *Social Ecology Working Paper, Vienna*, 100, 2008.

A. S. Drud. CONOPT – a large-scale GRG code. *ORSA Journal on Computing*, 6(2): 207–216, 1994.

R. E. Evenson. Spillover benefits of agricultural research: Evidence from U.S. experience. *American Journal of Agricultural Economics*, 71(2):447–452, 1989.

R. E. Evenson and D. Gollin. Assessing the impact of the green revolution, 1960 to 2000. *Science*, 300(5620):758–762, 2003.

F. Ewert, M. D. A. Rounsevell, I. Reginster, M. J. Metzger, and R. Leemans. Future scenarios of european agricultural land use:: I. estimating changes in crop productivity. *Agriculture, Ecosystems & Environment*, 107(2-3):101–116, 2005.

M. Fader, S. Rost, C. Müller, A. Bondeau, and D. Gerten. Virtual water content of temperate cereals and maize: Present and potential future patterns. *Journal of Hydrology*, 384(3-4):218–231, 2010.

S. Fan, L. Zhang, and X. Zhang. *Growth, inequality, and poverty in rural China: The role of public investments*. Int Food Policy Res Inst IFPRI, 2002.

FAOSTAT. Food & Agriculture Organization of the United Nations Statistics Division, accessed 8/11/2008, 2008. URL `http://faostat.fao.org`.

FAOSTAT. Food & Agriculture Organization of the United Nations Statistics Division, accessed 11/6/2009, 2009. URL `http://faostat.fao.org`.

FAOSTAT. Food & Agriculture Organization of the United Nations Statistics Division, accessed 11/2/2011, 2011. URL `http://faostat.fao.org`.

R. Färe, S. Grosskopf, M. Norris, and Z. Zhang. Productivity growth, technical progress, and efficiency change in industrialized countries. *The American Economic Review*, 84 (1):66–83, 1994.

R. Färe, S. Grosskopf, and D. Tyteca. An activity analysis model of the environmental performance of firms–application to fossil-fuel-fired electric utilities. *Ecological Economics*, 18(2):161–175, 1996.

G. Federico. *Feeding the world*. Princeton University Press, 2005.

L. E. Fulginiti, R. K. Perrin, and B. Yu. Institutions and agricultural productivity in Sub-Saharan Africa. *Agricultural Econmics*, 31:169–180, 2004.

R. Gentleman, B. Ding, S. Dudoit, and J. Ibrahim. Distance measures in DNA microarray data analysis. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 189–208, 2005.

D. Gerten, S. Schaphoff, U. Haberlandt, W. Lucht, and S. Sitch. Terrestrial vegetation and water balance–hydrological evaluation of a dynamic global vegetation model. *Journal of Hydrology*, 286(1-4):249–270, 2004.

D. Gerten, S. Schaphoff, and W. Lucht. Potential future changes in water limitations of the terrestrial biosphere. *Climatic Change*, 80(3):277–299, 2007.

C. C. Gibson, E. Ostrom, and T. K. Ahn. The concept of scale and the human dimensions of global change: a survey. *Ecological Economics*, 32(2):217–239, 2000.

M. Gosme, F. Suffert, and M. H. Jeuffroy. Intensive versus low-input cropping systems: What is the optimal partitioning of agricultural area in order to reduce pesticide use while maintaining productivity? *Agricultural Systems*, 103(2):110–116, 2010.

J. A. Hartigan. *Clustering algorithms*. Wiley New York, 1975.

J. A. Hartigan. Statistical theory in clustering. *Journal of classification*, 2(1):63–76, 1985.

L. D. D. Harvey. Upscaling in global change research. *Climatic Change*, 44(3):225–263, 2000.

T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27 (2):83–85, 2005.

M. Heistermann, C. Müller, and K. Ronneberger. Land in sight? achievements, deficits and potentials of continental to global scale land-use modeling. *Agriculture, Ecosystems & Environment*, 114(2-4):141–158, 2006.

M. Horridge and D. Laborde. TASTE a program to adapt detailed trade and tariff data to GTAP-related purposes. *GTAP Conference Paper*, 2666, 2008.

W. E. Huffman and R. E. Evenson. *Science for agriculture: A long-term perspective*. Wiley-Blackwell, 2006.

A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice Hall College Div, 1988.

H. Joe. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405):157–164, 1989.

B. F. Jones. The burden of knowledge and the 'Death of the renaissance man': Is innovation getting harder? *Review of Economic Studies*, 76(1):283–317, 2009.

R. W. Kates, G. Hydâen, and B. L. Turner II. Theory, evidence, study design. *Population growth and agricultural change in Africa*, pages 1–40, 1993.

S. Kellison. *The theory of interest*. McGraw-Hill/Irwin, 2nd edition, 1991.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

M.G. Kendall and J.D. Gibbons. *Rank Correlation Methods*. Oxford University Press, 5th edition, 1990.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 2002.

C.J. Kowalski. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Applied Statistics*, 20:1–12, 1972.

M. Krause, H. Lotze-Campen, and A. Popp. Spatially-explicit scenarios on global cropland expansion and available forest land in an integrated modelling framework. *27th International Conference of Agricultural Economists (IAAE), Bejing*, 2009.

D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo Methods*. Wiley, 1st edition, 2011.

J. Krüger. Hierarchie und skalen als ordnungsprinzipien für die abbildung des objektumfanges und der beschreibung kohärenter strukturen. *Leipziger Geowissenschaften*, 18:123–138, 2007.

E. F. Lambin, M. D. A. Rounsevell, and H. J. Geist. Are agricultural land-use models able to predict changes in land-use intensity? *Agriculture, Ecosystems & Environment*, 82(1-3):321–331, 2000.

H. Lotze-Campen, C. Müller, A. Bondeau, S. Rost, A. Popp, and W. Lucht. Global food demand, productivity growth and the scarcity of land and water resources: a spatially explicit mathematical programming approach. *Agricultural Economics*, 39 (3):325–338, 2008.

H. Lotze-Campen, A. Popp, T. Beringer, C. Müller, A. Bondeau, S. Rost, and W. Lucht. Scenarios of global bioenergy production: The trade-offs between agricultural expansion, intensification and trade. *Ecological Modelling*, 221(18):2188–2196, 2010.

T. Ma and Y. Nakamori. Modeling technological change in energy systems – from optimization to agent-based modeling. *Energy*, 34(7):873–879, 2009.

T. R. Malthus. *An essay on the principle of population*. Electronic Scholarly Publishing Project, 1998.

B. A. McCarl, A. Meeraus, P. van der Eijk, M. Bussieck, S. Dirkse, and P. Steacy. *McCarl GAMS User Guide*, 2008. URL `http://gams.com/docs/document.htm`.

C. Monfreda, N. Ramankutty, and J. A. Foley. Farming the planet: 2. geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochem. Cycles*, 22(GB1022):1–19, 2008.

J. Munksgaard, L. B. Christoffersen, H. Keiding, O. G. Pedersen, and T. S. Jensen. An environmental performance index for products reflecting damage costs. *Ecological Economics*, 64(1):119–130, 2007.

B. Narayanan and T. L. Walmsley. *Global Trade, Assistance, and Production: The GTAP 7 Data Base*. Center for Global Trade Analysis, Purdue University, 2008.

G. C. Nelson, M. W. Rosegrant, J. Koo, R. Robertson, T. Sulser, T. Zhu, C. Ringler, S. Msangi, A. Palazzo, M. Batka, M. Magalhaes, R. Valmonte-Santos, M. Ewing, and D. Lee. Climate change - impact on agriculture and costs of adaptation. *IFPRI Food Policy Report*, 21, 2009.

R.M.C. Netting. *Smallholders, householders: farm families and the ecology of intensive, sustainable agriculture*. Stanford University Press, 1993.

K. Neumann, P. H. Verburg, E. Stehfest, and C. Müller. The yield gap of global grain production: A spatial analysis. *Agricultural Systems*, 103(5):316–326, 2010.

A. Nin, C. Arndt, T. W. Hertel, and P. V. Preckel. Bridging the gap between partial and total factor productivity measures using directional distance functions. *American Journal of Agricultural Economics*, 85(4):928–942, 2003.

M. Nishimizu and J. M. Page. Total factor productivity growth, technological progress and technical efficiency change: Dimensions of productivity change in Yugoslavia, 1965-78. *The Economic Journal*, 92(368):920–936, 1982.

OECD. International transport forum - infrastructure investment and maintenance database, accessed 26/11/2010, 2010. URL `http://www.internationaltransportforum.org/statistics/investment/data.html`.

OECD-FAO. Agricultural outlook 2009-2018, 2009. URL `http://www.agri-outlook.org`.

F. Oehl, E. Sieverding, K. Ineichen, P. Mader, T. Boller, and A. Wiemken. Impact of land use intensity on the species diversity of arbuscular mycorrhizal fungi in agroecosystems of central europe. *Applied and Environmental Microbiology*, 69(5):2816–2824, 2003.

P. G. Pardey and N. Beintema. *Slow Magic - Agricultural R&D a Century After Mendel*. Food Policy Report: International Food Policy Research Institute, 2001.

P. G. Pardey and B. Craig. Causal relationships between public sector agricultural research expenditures and output. *American Journal of Agricultural Economics*, 71 (1):9–19, 1989.

P. G. Pardey, N. Beintema, S. Dehmer, and S. Wood. *Agricultural Research: A growing global divide?* Agricultural Science and Technology Indicators Initiative: International Food Policy Research Institute IFPRI, 2006.

A. Popp, H. Lotze-Campen, and B. Bodirsky. Food consumption, diet shifts and associated non-CO2 greenhouse gases from agricultural production. *Global Environmental Change*, 20(3):451–462, 2010.

F. Portmann, S. Siebert, and P. Döll. Mirca2000 - global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochem. Cycles*, 24(GB1011):1–24, 2010.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`.

S. Reinhard, C. A. Lovell, and G. Thijssen. Econometric estimation of technical and environmental efficiency: An application to dutch dairy farms. *American Journal of Agricultural Economics*, 81(1):44–60, 1999.

J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

P. M. Romer. Endogenous technological change. *The Journal of Political Economy*, 98 (5):71–102, 1990.

J. Roseboom. Underinvestment in agricultural R&D revisited. *Quarterly Journal of International Agriculture*, 41(4):297–316, 2002.

M. W. Rosegrant, C. Ringler, S. Msangi, T. B. Sulser, T. Zhu, and S. A. Cline. *International model for policy analysis of agricultural commodities and trade (IMPACT): Model description*. International Food Policy Research Institute (IFPRI), 2008.

M. R. Rosenzweig, H. P. Binswanger, and J. McIntire. *From land abundance to land scarcity - The effects of population growth on production relations in agrarian economies*. Clarendon Press, 1988.

S. Rost, D. Gerten, A. Bondeau, W. Lucht, J. Rohwer, and S. Schaphoff. Agricultural green and blue water consumption and its influence on the global water system. *Water Resources Research*, 44(W09405):1–17, 2008.

C. F. Runge, B. Senauer, P. G. Pardey, and M. W. Rosegrant. *Ending Hunger in our Lifetime - Food Security and Globalization*. The John Hopkins University Press, 2003.

V. W. Ruttan. Bureaucratic productivity: the case of agricultural research. *Public Choice*, 35(5):529–547, 1980.

R. D Sands and M. Leimbach. Modeling agriculture and land use in an integrated assessment framework. *Climatic Change*, 56(1):185–210, 2003.

U. A. Schneider and D. E. Schwab. The european forest and agricultural sector optimization model. In *Proceedings, IATRC Annual Meeting*, pages 4–6, 2005.

R. W. Shephard. *Theory of cost and production functions*. Princeton University Press, 1970.

A. J. Shriar. Agricultural intensity and its measurement in frontier regions. *Agroforestry Systems*, 49(3):301–318, 2000.

S. Sitch, B. Smith, I. C Prentice, A. Arneth, A. Bondeau, W. Cramer, J. O Kaplan, S. Levis, W. Lucht, M. T. Sykes, et al. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, 9(2):161–185, 2003.

R. A. Slaughter. Long-term thinking and the politics of reconceptualization. *Futures*, 28(1):75–86, 1996.

M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526, 2000.

N. Stephenne and E. F. Lambin. A dynamic simulation model of land-use changes in sudano-sahelian countries of africa (SALU). *Agriculture, Ecosystems & Environment*, 85(1–3):145–161, 2001.

J. Stoer and R. Bulirsch. *Introduction to numerical analysis*. Springer, 2002.

C. Thirtle, L. Lin, and J. Piesse. The impact of research-led agricultural productivity growth on poverty reduction in africa, asia and latin america. *World Development*, 31 (12):1959–1975, 2003.

A. Trewavas. Malthus foiled again and again. *Nature*, 418(6898):668–670, 2002.

B. L. Turner and W. E. Doolittle. The concept and measure of agricultural intensity. *The Professional Geographer*, 30(3):297–301, 1978.

M. G. Turner, R. H. Gardner, and R. V. O'Neill. *Landscape ecology in theory and practice: pattern and process*. Springer Verlag, 2001.

United Nations. World population prospects, the 2004 revision, 2005. URL `http://www.un.org/esa/population/publications/sixbillion/sixbilpart1.pdf`.

M. K. van Ittersum and R. Rabbinge. Concepts in production ecology for analysis and quantification of agricultural input-output combinations. *Field Crops Research*, 52(3): 197–208, 1997.

H. van Meijl and F. van Tongeren. Endogenous international technology spillovers and biased technical change in agriculture. *Economic Systems Research*, 11(1):31–48, 1999.

G. van Rossum and F. L. Drake Jr. *Python reference manual*, 2001. URL `http://docs.python.org/release/2.2/ref/ref.html`.

P. H. Verburg, G. H. J. De Koning, K. Kok, A. Veldkamp, and J. Bouma. A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use. *Ecological modelling*, 116(1):45–61, 1999a.

P. H. Verburg, A. Veldkamp, and L. O. Fresco. Simulation of changes in the spatial pattern of land use in china. *Applied Geography*, 19(3):211–233, 1999b.

R. Verburg, E. Stehfest, G. Woltjer, and B. Eickhout. The effect of agricultural trade liberalisation on land-use related greenhouse gas emissions. *Global Environmental Change*, 19(4):434–446, 2009.

K. Waha, L. G. J. van Bussel, C. Müller, and A. Bondeau. Climate-driven simulation of global crop sowing dates. *Global Ecology and Biogeography*, 2011.

T. Wassenaar, P. Gerber, P. H. Verburg, M. Rosales, M. Ibrahim, and H. Steinfeld. Projecting land use changes in the neotropics: the geography of pasture expansion into forest. *Global Environmental Change*, 17(1):86–104, 2007.

C. A. Wessman. Spatial scales and global change: Bridging the gap from plots to GCM grid cells. *Annual Review of Ecology and Systematics*, 23(1):175–200, 1992.

M. Wik, P. Pingali, and S. Broca. Global agricultural performance: past trends and future prospects. *Background paper for the World Development Report 2008*, 2008.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6): 80–83, 1945.

World Bank. World development indicators (CD-ROM), Washington DC, 2001.

H. G. Zechmeister and D. Moser. The influence of agricultural land-use intensity on bryophyte species richness. *Biodiversity and Conservation*, 10(10):1609–1625, 2001.

List of Figures

List of Tables

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Potsdam, Mai 2011

Jan Philipp Dietrich