From the Institute of Natural Science

at the Department of Landscape, Water and Biogeochemical Cycles

of the Justus-Liebig University Gießen

Uncertainty analysis of complex hydro-biogeochemical models

Dissertation from M.Sc. Tobias Houska

for the degree Doctor of Natural Science (Dr. rer. nat.)

Referees from the Justus-Liebig-University Gießen:

Prof. Dr. Lutz Breuer	(1st Supervisor)
Prof. Dr. Christoph Müller	(2nd Supervisor)
Prof. Dr. Jürg Luterbacher	
Prof. Dr. Jan Siemens	

Submitted: 14th June 2017

Abstract

This thesis is about complex hydro-biogeochemical models and their practical applications. Several modelling practices and their associated uncertainty are investigated in this joined project of the working groups of Prof. Dr. Lutz Breuer, Justus Liebig University Giessen, and Prof. Dr. Klaus Butterbach-Bahl at Karlsruhe Institute of Technology. The aim of the project is to develop strategies for reducing the climate footprint of agricultural production and to quantify uncertainties of model-based strategies for low emission pathways, while at the same time increasing the credibility in model predictions by evaluating not only trace gas emissions, but also plant growth and hydrological fluxes. A motivation that is next to me driven by an increasing demand of the scientific community, governmental and non-governmental organizations.

During the three-year's project, I setup different methods to access parameter sensitivity, parameter and structure uncertainty of environmental models. The methods were combined in a statistical parameter optimization tool for python (SPOTPY) to perform various model diagnostics in a straightforward way. Both working groups and others use the tool now in joined as well as individual studies.

The SPOTPY package enabled us to gain a deeper understanding of the underlying processes and limitations of the investigated complex hydro-biogeochemical models. A key result of my study is that the tested models still lack on robustness to generate outputs for multiple ecosystem services. A change of awareness of site and data managers is required as sensors and small-scale variability of site properties can cause low performance in terms of model predicting capability.

Under this impression, I equipped a study area with greenhouse gas emissions and other important water, carbon and nitrogen fluxes measurements on arable land, grassland and forest. I used these measurements in a model-data fusion approach as a final contribution to my dissertation. This study allowed me to derive missing model processes that would potentially increase model simulation performances, if implemented into the biogeochemical model. My findings provide a strong motivation to enhance our understanding of the hydro-biogeochemical system and guide future work.

Table of contents

Extended Summary

Water, carbon (C) and nitrogen (N) are key elements in all ecosystems and turnover processes within them. They are related to a variety of environmental problems, including droughts and floods (for water), eutrophication, drinking water quality, fish toxicity, soil N_2O and NO emissions (for nitrogen) or C sequestration, CO_2 and CH_4 emissions (for carbon). Given this, an in-depth knowledge of the interaction of water, C and N on the landscape scale is required to improve land use and management while at the same time mitigating environmental impact. Cultivated landscapes are affected by a multitude of such managements, e.g. fertilization, grazing or deforestation. Consequently, not only the carbon/nitrogen pools and the microbial communities change. The underlying processes like denitrification, nitrification and respiration react immediately on changes in available nutrients. Consequently, high variability of greenhouse gas (GHG) emissions across space and time are reported (McClain et al., 2003). This variability across the spatio-temporal scales cannot be addressed by field measurement only as the spatial scale is too limited (Butterbach-Bahl et al., 2013).

To overcome the current measurement limitations to quantify processes underlying the biosphere-atmosphere GHG interaction, biogeochemical models have been developed, which translate our current understanding into numerical equations. These models allow upscaling in space and time domains to estimate GHG emissions, where no measurements exists. The individual modeling of hydrological and biogeochemical fluxes on the landscape is well represented in the literature. A variety of different model approaches spanning from empirical-conceptual to process-oriented methods are available: Reaching from low complexity but fast model runtime with the model CENTURY (Parton et al., 1988), which is driven by lumped parameters for C, N, P and S fluxes to high temporal resolution with the updated version DAYCENT (Parton et al., 1998). Medium complexity is given with models like CERES-EGC (Gabrielle et al., 2006) a process-based biogeochemical extension of the CERES crop model. The most complex process-based coupled hydrological and biogeochemical model are RHESSys (Tague and Band, 2004) and LandscapeDNDC (Haas et al., 2013), both covering different spatial and temporal resolutions.

Biogeochemical model application studies have been published for site, regional, continental or global scales where the model runs separately at one or a large number of grid cells (Dai et al., 2012; Li et al., 2004; Rosenzweig et al., 2014; Werner et al., 2007). These studies aim to upscale site scale applications to get inventories of GHG emissions – however, they remain one-dimensional, neglecting potentially important horizontal fluxes of water and nutrients. Therefore, Haas et al. (2013) developed a framework to facilitate regional applications of biogeochemical models to overcome this

limitation. In contrast to other models, all cells are synchronized in time. This is not only important for upscaling, but also highly significant for model-independent communication when it comes to coupling. However, complex process-based and/or coupled models are, like all environmental models, prone to uncertainty with regard to their parameterization, structure and input data. There is an intensive discussion about different sorts of model uncertainties and how to address them (Beven, 2015). Given the large number of parameters particularly in the LandscapeDNDC model, my thesis focuses on methods and results related to the parameterization and setup of the coupled LandscapeDNDC-CMF model:

Parameter uncertainty – the point of origin

Various methods are available to access parameter uncertainty. They all follow some general steps, which I also used in this project: In a first step, a selection of sensitive model input parameters has to be defined. This can be done by expert knowledge or through a sensitivity analysis. The more sensitive a model parameter for predicting a given target value is, the more it gets constrained through a parameter uncertainty analysis. Naturally, the efficiency of simulations decreases with the number of parameters. In a next step, the user has to define *a priori* distribution of every parameter in the analysis. The prior distribution comprises the knowledge a user has about the parameter. If no prior knowledge is given, a uniform distribution can be assumed, bounded by the physically possible settings of each parameter. The parameter are then altered through random (e.g. Monte Carlo) or stratified (i.e. depending on the model results, e.g. DREAM (Vrugt et al., 2009)) sampling, where the model is executed for each parameter realisation. The performance of a parameter set driving the model to predict observations is then evaluated by a "goodness-of-fit" value, represented by an objective function (depending on the research topic often also termed as likelihood, cost or signature function). The choice of the function depends on the situation and is often subjective, if no accurate information about the probability distribution of the measurement errors is available (Beven and Binley, 1992). The choice of only one function to access the parameter uncertainty is in most cases inaccurate (Vrugt et al., 2003) and has a strong influence on the results (He et al., 2010). Popular objective functions are, e.g. the Nash and Sutcliffe model efficiency (Freer et al., 1996), the inverse error variance with a shaping factor (Beven and Binley, 1992), scaled maximum absolute residuals (Keesman and van Straten, 1990) as well as the index of agreement (Wilmott, 1981), model bias and coefficient of determination. Thresholds (also known as limits of acceptability) of selected objective functions are then used to group model realizations into behavioural and non-behavioural. The former describes an acceptable model application, allowing some degree of error in simulating a target value (defined in an *a priori* threshold criteria). The latter describes parameter sets which return

unacceptable model outputs and can be deleted (Beven, 2006). The associated parameter sets to the behavioural model runs are defined as *posterior* parameter distribution, which can be interpreted with its range as the parameter uncertainty of a given model. A further distinction is made between constrained and unconstrained parameters (Christiaens and Feyen, 2002). The more sensitive a model parameter for predicting a given model output is, the more it gets constrained in the remaining *posterior* parameter sets.

One suitable method to screen the hyper-dimensional parameter space for the underlying uncertainty is the GLUE (Generalized Likelihood Uncertainty Estimation) method. GLUE is a widespread Bayesian technique and follows the above-defined general steps to investigate the parameter uncertainty (Beven and Binley, 1992). Since the establishment of GLUE for hydrological model applications in the 1990ies, a large number of studies used the method to gain a better understanding of the model performance and their input parameters. Nowadays, these applications are not restricted to hydrological modelling, but cover many fields of ecology. For example, Wang et al. (2005) utilized the GLUE method for evaluation of the EPIC plant growth model with the mean squared error as an objective function. Mo and Beven (2004) applied the method with the index of agreement as an objective function for calibration of a soil-vegetation-atmosphere-transfer model. During the past 10 years, first studies addressed the underlying parameter uncertainty of very complex biogeochemical models, i.e. a subgroup of environmental models, with Bayesian techniques (Del Grosso et al., 2010; van Oijen et al., 2011). Since then, a limited number of biogeochemical model studies extended the application to uncertainty analysis of GHG exchange processes and fluxes. They differ with respect to techniques used to access uncertainty, implemented process descriptions, output targets. Wang and Chen (2012) summarized the few existing uncertainty analysis in the biogeochemical community. However, under the viewpoint of model improvement, parameter uncertainty analysis is not the answer to everything, as it does not give information about model structural deficiencies.

Going beyond parameter uncertainty – uncertain model structures

Recently, two similar calls were made in the hydrological community by Clark et al. (2011) and in the biogeochemical community by Wang and Chen (2012). Both recommend uncertainty analysis for multiple sites and the use of multiple criteria. They further suggest a development of a model library containing various model structures to facilitate comprehensive model comparison and uncertainty studies. So far, such variable model structures are only available for a very limited number of modelling frameworks.

In the biogeochemical community LandscapeDNDC (DeNitrification-DeComposition) is one such framework with a variable model structure (Haas et al., 2013):

> LandscapeDNDC is a modelling framework for the simulation of water, C and N cycling and associated GHG emissions in terrestrial (forest, arable, grassland) ecosystems. LandscapeDNDC consists of interchangeable modules representing soil biogeochemistry (e.g., scDNDC (Zhang et al., 2015) or MeTr^x (Kraus et al., 2015)) or soil hydrology (e.g., wcDNDC) as well as various modules for vegetation and microclimate processes. A setup of LandscapeDNDC is done by writing different xml files, which contain information about meteorology, soil and management. A large number (n > 130) of input parameters are needed for each of the different modules. In this thesis, I used LandscapeDNDC to simulate GHG emissions and the C and N cycle and reduced the number of parameters through a sensitivity analysis to n = 30.

For the hydrological community Kraft et al. (2011) developed the Catchment Modelling Framework (CMF) with the possibility to build a hydrological model with pre-build process implementations:

> CMF is a computer program to setup individual hydrological models, following the finite volume approach. A programming library facilitates the design of water transport models between soil layers in up to three-dimensions. A network of storages defines models in CMF and boundary conditions connected equations calculate the flux between them. It allows the development of detailed mechanistic models as well as lumped large-scale linear storage-based models, ranging from simple linear water flux connections (e.g. Kinematic wave) to complex nonlinear partial differential functions (e.g. Richards equation). A model build with CMF functions as a network of storages and boundary conditions connected by flux-calculating sub-models. It works as an extension to Python and is connectable with other models, as realized for example by Haas et al. (2013) and Houska et al. (2014).

Both frameworks have proven their general potential to reproduce observed data in various publications (e.g. Houska et al., 2014; Molina-Herrera et al., 2016; Windhorst et al., 2014; Zhang et al., 2015). However, to achieve a reliable simulation of GHG emissions, an accurate representation of the soil moisture is a key requirement (Butterbach-Bahl et al., 2013). Nevertheless, in biogeochemical models, soil hydrological processes are often simulated based on simple bucket approaches, i.e. water moves vertically down a profile once a certain threshold has been reached, as e.g. in the LandscapeDNDC hydrological module wcDNDC. The nonlinear Richards' equation brings

the advantage of a complex physical based approach. The equation describes vertical unsaturated flow, capillary rise and interaction with groundwater level. Such water fluxes are particularly important in lowland, groundwater impacted ecosystems such as meadow and wetlands or more generally, the riparian zone. Therefore, a coupling of both frameworks was performed, to improve model performance:

> LandscapeDNDC-CMF is a coupled approach to enhance process based modelling and enabling the application of three-dimensional setups. Model coupling was either achieved using a MPI-based PALM coupler (Wlotzka et al. 2014) or by CMF model integration into LandscapeDNDC (Klatt et al., 2017). The latter approach was used in my dissertation. During simulation, biogeochemical and hydrological models receive continuous climatic inputs. The modelling focus is on water fluxes either from a cell to its neighbours, to outlets or within the cell's soil layers. These water fluxes are modelled based on the Richards equation. CMF provides access to these flux values at every time step allowing the estimation the amount of transported solutes. The model was tested with a virtual hillslope (Haas et al., 2013), with a virtual landscape (Wlotzka et al., 2014) and with consideration of vegetated buffer strips in a virtual landscape (Klatt et al., 2017).

In my dissertation, I aim to quantify the underlying model structure uncertainty of the presented hydro-biogeochemical frameworks. In particular, I am interested in the benefits of a physically based process description over a conceptual approach in simulating soil water dynamics within a biogeochemical model. I follow the philosophy that complex models should be identifiable (low parameter uncertainty) and accurate (good agreement with observation data). Further, a model should be able to simulate various observation data concurrently and close to reality, especially when dealing with highly non-linear process interactions like in hydro-biogeochemical systems. To asses only such model runs, I perform a multi-criteria evaluation of different model structures and quantify their underlying uncertainties. In order to achieve meaningful results, I require comprehensive observation data, complex process based models and powerful tools to analyse the results. During the three-year project, I worked on these points and came up with a meaningful uncertainty analysis of the hydro-biogeochemical frameworks. In a first study, I tested different uncertainty estimation techniques and objective functions (chapter I), to access parameter uncertainty (chapter II) and model structure uncertainty (chapter III) by using multiple objective functions for multiple model outputs of the different frameworks.

New tool for model parameter optimization

A wide variety of different methods is available to access model parameter settings. In order to make the test of methods in a straightforward way possible, I developed SPOTPY as an open source Python library. As first part of my thesis, I enable with this tool the use of computational optimization techniques for calibration, uncertainty and sensitivity analysis techniques on almost every (environmental-) model with high-performance computer cluster support (see chapter I):

Houska, T., Kraft, P., Chamorro-Chavez, A. and Breuer, L.: SPOTting Model Parameters Using a Ready-Made Python Package, PLoS ONE, 10(12), e0145180, doi:10.1371/journal.pone.0145180, 2015.

General functionality

The package comprises thirteen widely used algorithms for uncertainty analysis, optimization and sensitivity analysis (Table 1) and thirteen different objective functions. SPOTPY supports to test and use different setups of parameter estimation methods and makes the application on high performance computing clusters possible. All algorithms realized in the SPOTPY package can work with build-in parameter distributions and objective functions, which allows their use for multi-objective calibration approaches. A progress bar enables to monitor the sampling. The use of highly optimized python code makes the time needed for the parameter sampling, the model starting and the results saving short. Two different databases solutions are currently available: *ram* storage for fast sampling and *csv* tables, the secure solution for long duration samplings. After sampling, the best run is returned together with its underlying parameter setting. A build-in analyser is designed to plot parameter traces, parameter interaction (including the Gaussian-kde function), regression analysis between simulation and evaluation data, posterior parameter distribution and convergence diagnostics (including Gelman-Rubin and Geweke statistics). To setup a model with SPOTPY, the tool comes along with a wide range of pre-build coding examples and tutorials.

Selection of objective function

In a first case study, I used outputs of the biogeochemical model LandscapeDNDC and compared them with measured CO_2 emissions data of a long-term grassland study site. The emissions were measured with the dark closed chamber method (Kammann et al., 2001). I chose different objective functions from the SPOTPY package to quantify the goodness-of-fit and run a Latin Hypercube based calibration with n =50,000 model runs. Depending on the objective function best model runs were selected and compared. With this, the tool allowed a fast application of a case study, which can be used to gain background knowledge, e.g. that model performance can be flawed when simulations are analysed with an inappropriate objective function. For example, the objective function BIAS is suited to reduce the mean error, but it does not guarantee that the model fits the temporal dynamic of

the evaluation data. The coefficient of determination, also known as r^2, is suited to find parameter sets driving the underlying model to fit the timing of the system, but this objective function does not guard against a systematic over or underestimation of the evaluation data. Legates and McCabe (1999) pointed out that the coefficient of determination is improper for model quantification because it is oversensitive to high flow but insensitive to additive and proportional differences between model simulations and observations. They recommended RMSE as the model evaluation tools. We could show that the root mean squared error (RMSE) and the agreement index (AI) are well suited to find model realizations fitting the absolute values of the observed data. A more general classification of different objective functions into two types was done by Guinot et al. (2011): distance-based objective function (e.g. RMSE) and weak form-based objective function (e.g. BIAS and r^2). They concluded that although the distance-based objective functions have the advantage to search an identifiable model-parameter set, they might have problems caused by local extremes in the response surface and lead to mis-calibration, i.e. being trapped around local optima. By contrast, the weak form-based objective functions are more monotone than the distance-based objective functions. Depending on the objective of the model approach, it can be beneficial to combine several objective functions to find reliable posterior simulations. While this is not a surprising or new result, the advantage of SPOTPY is, that it facilitates an easy comparison of objective functions in a pre- and post-processing mode.

Table 1. Available algorithms implemented in SPOTPY. Given are the acronyms, the number of citations of the corresponding publication (based on Google Scholar search results in May 2017), the full name and the authors of the corresponding publication.

Model diagnostic	**Non-Bayesian Calibration**	**Bayesian Calibration**
FAST - 210 citations *Fourier Amplitude Sensitivity Test* McRae et al. (1982)	**SCE-UA** – 2,694 citations *Shuffled Complex Evolution Uncertainty Analysis* Duan et al. (1992)	**DREAM** - 277 citations *Differential Evolution Adaptive Metropolis* Vrugt et al. (2009)
GLUE - 2,785 citations *Generalized Likelihood Uncertainty Estimation* Beven and Binley (1992)	**FSCABC** - 231 citations F*itness Scaled Chaotic Artificial Bee Colony* Zhang et al. (2011)	**DE-MCz** - 90 citations *Differential Evolution Markov Chain* terBraak and Vrugt (2008)
	ABC – 2,676 citations *Artificial Bee Colony* Karaboga and Basturk (2007)	**MCMC** - 29,082 citations *Markov Chain Monte Carlo* Metropolis et al. (1953)
	MLE - 10,766 citations *Maximum-Likelihood Estimation* e.g. Johansen (1990)	**LHS** - 4,938 citations *Latin Hypercube Sampling* McKay (1979)
	SA - 32,763 citations *Simulated Annealing* Kirkpatrick and Vecchi (1983)	**ROPE** - 67 citations *Robust Parameter Estimation* Bardossy and Singh (2008)
		MC -2,528 citations *Monte Carlo* e.g. Fishman (1996)

Selection of parameter estimation methods

Wagener and Gupta (2005) reviewed different uncertainty estimation methods. They conclude that these methods differ in the underlying philosophy, assumptions and sampling strategies. They found remarkably poor understanding of the effect of these differences, and woefully little guidance on which approach should be used under what circumstances to analyse models and their simulations. In order to increase the understanding of differences, I tested the effect of different parameter estimation methods (Table 1) and give a general guidance, which algorithm is suitable for which application (Houska et al., 2015). The case studies included three different numerical optimization problems and one hydrological model application. Results showed that every algorithm had its strengths in particular parameter search problems. Inspired by the results, I developed a decision-tree to help possible users to choose one of the implemented algorithms (Figure 1). One of the used algorithms (Differential Evolution Adaptive Metropolis; DREAM) was included during a cooperation visit of mine to the University of California, Irvine, together with Prof. Dr. Jasper Vrugt (Figure 2).

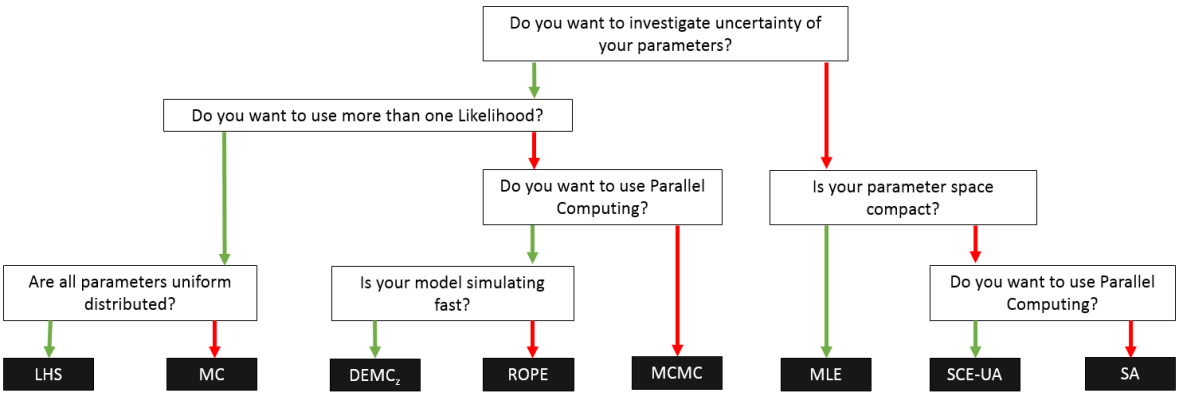

Figure 1. Decision-tree as a guidance for the choice of an algorithm in SPOTPY for a specific optimization problem.

Comparison to other packages

A surprisingly small range of software applications is available giving users' access to tests different parameter estimation methods. One of them is PEST, a GUI program for Model-Independent Parameter Estimation and Uncertainty Analysis. Others are OpenBugs and Jags (for performing Bayesian inference Using Gibbs Sampling), PyMC (Comprehensive Python package to analyse models with Marcov Chain Monte Carlo (MCMC) techniques), STAN (implementing MCMC techniques like NUTS, HMC and L-BFGS), emcee (Affine Invariant MCMC Ensemble sampler) and BIP (Bayesian inference with a DREAM sampler). All of them have their pros and cons which are outlined in chapter III. None of the packages can offer a very wide range of different algorithms. To perform a benchmark of multiple algorithms against each other, as I have done in this project, a comprehensive and error prone combination of these packages would be required.

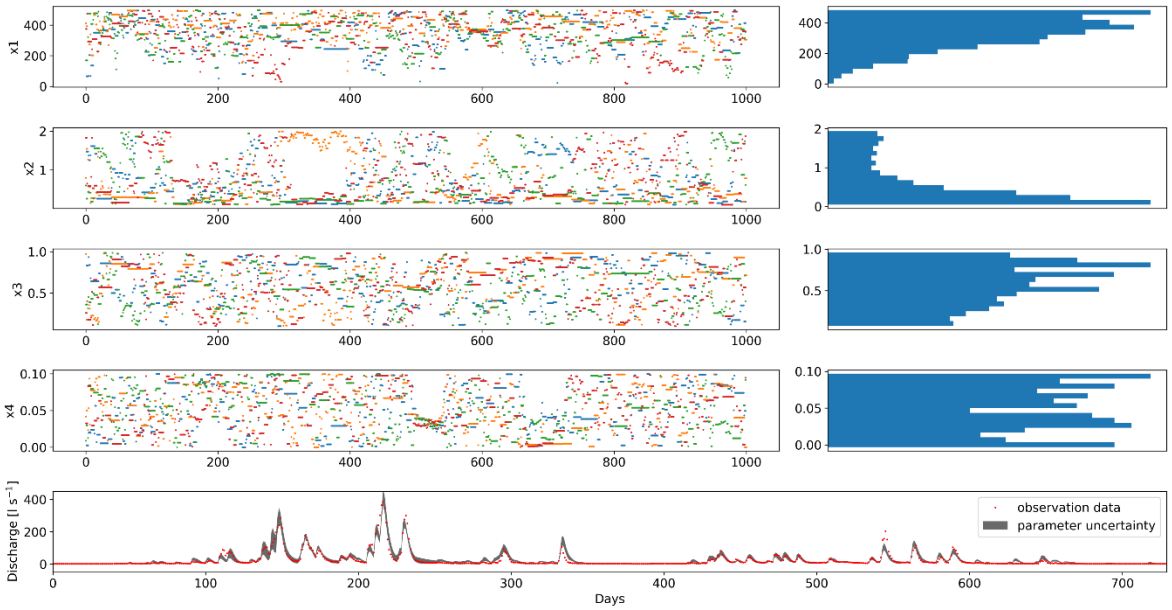

Figure 2. Example output of SPOTPY using the DREAM algorithm. Parameter traces can be plotted for every Metropolis chain (different colored dots for four parameters x1-x4, upper left panels), together with their probability distribution (bar plots, upper right panels) and remaining posterior model runs (dark grey uncertainty band around observations, bottom panel). Observation discharge data is given as red dots.

Uncertainty analysis LandscapeDNDC-CMF

In a second contribution to my dissertation, I tested SPOTPY within a complex modelling approach. For this, I selected the biogeochemical model framework LandscapeDNDC and the hydrological model framework CMF, as described in chapter II:

> Houska, T., Kraft, P., Liebermann, R., Klatt, S., Kraus, D., Haas, E., Santabárbara, I., Kiese, R., Butterbach-Bahl, K., Müller, C. and Breuer, L.: Rejecting hydro-biogeochemical model structures by multi-criteria evaluation, Environ. Model. Softw., 93, 1–12, doi:10.1016/j.envsoft.2017.03.005, 2017.

My study builds on a 2011 paper in *Environmental Modelling and Software* by Kraft and colleagues (Kraft et al. 2011) that presented CMF to build hydrological models. CMF was included into the LandscapeDNDC model by Haas et al. (2013) with the goal to improve hydrological process description in biogeochemical modelling. Coupling of both frameworks, required structural changes in the code of LandscapeDNDC and the establishment of an effective communication structure between both models for exchanging state conditions (e.g. soil moisture, nutrient loading, soil solute concentration, thermal conditions). The general challenges of code adaptation, modernization and coupling were mainly addressed within associated DFG funded projects (BU 1173/12-1/; HE 4760/4-1). Model coupling was either achieved using a MPI-based PALM coupler (Wlotzka et al., 2014) or by CMF model integration into LandscapeDNDC (Klatt et al., 2017). The latter approach was used in my project to test the benefits of this coupled model with real world data:

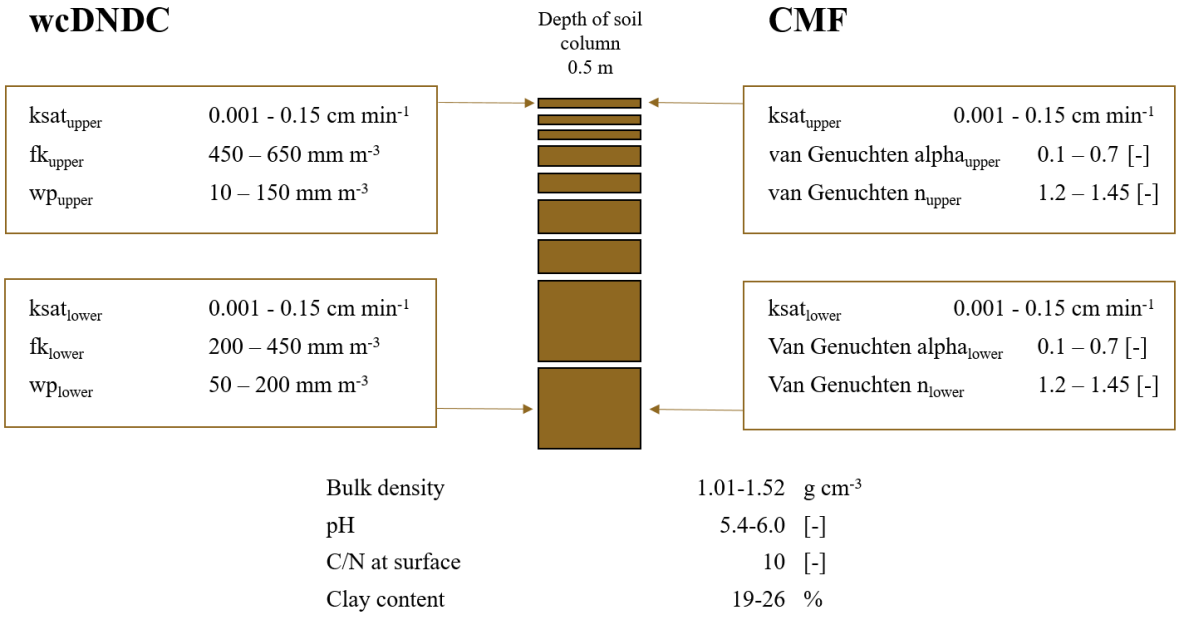

Figure 3. Differences in the soil setup of the hydrological modules wcDNDC and CMF inside the LandscapeDNDC framework. Parameter boundaries are given for each module for the highest and the lowest soil layer. Parameter settings in between are derived by a depth function based on linear regression. Fk = field capacity, wp = wilting point, ksat = saturated conductivity.

The model uncertainty was analysed with long-term (8 years) data on management, soil GHG emissions and water filled pore space at a meadow wet grassland site in Linden (Germany). The study site consists of six plots, of which three are treated with elevated with CO_2 (E1-3) and three remain under ambient conditions (A1-3). In this second already published study of my thesis, I only used the ambient plots. The study site has a mean annual precipitation amount of 616 mm and an average annual temperature 9.5°C. The vegetation is characterized as grassland defined by the main species *Arrhenatheretum elatioris* and the soil is a fluvic gleysol (Kammann et al., 2008).

For model validation, we used measured GHG emissions of each plot, which are determined on a weekly basis with dark closed static chambers (0.3 m height, 0.184 m^3 volume). Chambers were sealed for 60-90 min to a soil collar and sampled in four 20-30 min time intervals (Kammann et al., 2008). Samples were analysed within 24 h for CO_2 and N_2O content with a gas chromatograph (HP6890) and GHG fluxes are calculated according to Kammann et al. (2008) from the linear increase of GHG concentrations within the chambers. Corresponding to the dark chamber measurement method, where a lightproof chamber is placed over the plants on the soil (see Kammann et al. (2008) for details), the measured soil CO_2 emissions were compared to the sum of simulated heterotrophic and autotrophic maintenance respiration of the plants. They reflect the respiration of the soil. Simulated autotrophic growth respiration was excluded, assuming that photosynthesis stops with chamber closure. This provides a way to model the measured plant-physiology darkness. Volumetric soil moisture of each plot is measured on working days with TDR sensors in 0-10 cm depth (Kammann et al., 2008). The vegetation is harvested in June and September each year 4 cm above the soil surface and fertilized in April with 40 kg N ha^{-1} a^{-1} consisting of granular mineral calcium-ammonium nitrate (Kammann et al., 2008).

In order to perform a model structure uncertainty analysis, I choose four different model structures: two different biogeochemical modules of LandscapeDNDC (i.e. the widely used scDNDC and the newly developed $MeTr^x$ (Kraus et al., 2015)) and two hydrological modules (i.e. the simple wcDNDC and a complex soil moisture routines realized by CMF, compare Figure 3). They were used to reproduce long-term measured observation data of soil moisture, soil respiration, N_2O flux and biomass yield. I applied a sensitivity analysis (Fourier amplitude sensitivity test, FAST) in a first step, to reduce the parameters of the biogeochemical model from 130 to 30 (Houska et al., 2017b). In a second step, I tested with the remaining parameters in a new developed rejectionist framework. With rejection, I mean the selection of only those model structures that meet predefined objective functions thresholds to gain the posterior distribution. A Latin Hypercube sampling with the GLUE method was performed with resulting model runs evaluated by 84 different evaluation criteria. For the

selection of the posterior model runs, I make use of RMSE and the BIAS. I applied those objective functions for different outputs, measurement sites and observation years. To accept model runs as behavioral, I set strict limits based on the measured data and a literature review.
The results show that only 0.01% of all model runs (n = 400,000) passed the complete rejectionist framework (Figure 4). Here, I provide evidence that each model combination had its strength for particular criteria and that hardly any combination fulfilled the complete set of the 84 criteria. Regarding to the model intercomparison $MeTr^x$/CMF was better in simulating soil moisture, $MeTr^x$/wcDNDC was better in simulating CO_2 emissions and scDNDC/wcDNDC was better in simulating N_2O emissions. These results will guide the module selection for future model applications.

I compared my findings with other biogeochemical studies and my findings reveal that modelling efficiency dramatically drops from 40 to 70% (for frequently published single evaluation criteria) down to 0.01% when multiple evaluation criteria are used. My study indicates that models can be right for the wrong reasons, i.e., matching GHG emissions while at the same time failing to simulate other criteria such as soil moisture or plant biomass dynamics. These results indicates that care has to be taken to avoid that models matching GHG emissions at the same time fail to accurately meet other criteria such as a realistic representation of the water filled pore space or the growth of biomass. Unfortunately, no matter how good environmental models are setup and run with measured forcing data (e.g. soil information, fertilizer application and climate data), model parameter and structural uncertainties are likely to be misleading. I recommend that complex, process-based hydro-biogeochemical models need to be thoroughly tested and checked against multiple-criteria using appropriate objective functions.

Despite the questionable efficiency of the model structures to represent all observed data sets at the same time, the observation data itself are highly variable. The data measured on three plots at the grassland study site in Linden are supposed to have the same land use and soil type within a distance of less than 100 m on even topography. However, the daily measured water filled pores space (37.2±9.0, 46.2±11.6, 40.1±11.2 vol.%) and the weekly measured CO_2 emissions (25.0±75.4, 26.1±78.3, 23.8±67.8 kg C ha^{-1} day^{-1}) data show significant differences between the plots. Assuming homogenous soil for the three plots the only other possible impact influencing the measurements is the groundwater table. We cannot quantify the difference in the groundwater table yet, nor do we know the dynamic of possible N supply through upwelling groundwater. However, it might explain at least a part of the remaining model errors, as the groundwater reaches heights of 0.1 m below surface throughout the year and is in its flow path influenced by agriculture. Based on the model

results, detailed measurements of groundwater table depths (subdaily) and groundwater quality (total nitrogen and nitrate concentration) have been started to investigate the dynamic and estimate potential nitrogen input through upwelling groundwater.

Nevertheless, I face several question about data uncertainty and data quality: Are the available data sets sufficient for model evaluation? How do we deal with data uncertainty in larger scales? How can we improve the measurements? Are weekly GHG measurements sufficient to capture all relevant processes? Future modelling studies can help to address these questions, by having a closer look on the process validation instead of just fitting models on observed data. Concluding, with this study I provide a new developed method to perform biogeochemical model structures uncertainty analysis. I was able to show that process based hydrological modelling can improve biogeochemical model predictions. Moreover, my study raises awareness that modelling efficiency dramatically drops with multiple objectives.

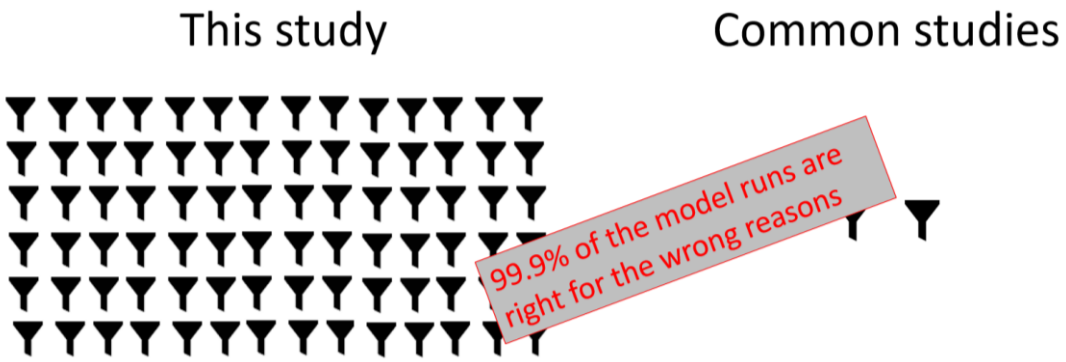

Figure 4. Difference in selection of 84 objective function thresholds based on RMSE and BIAS for different sites, target model output values and analysis in time vs. common studies with 1-2 thresholds.

Model-data fusion with LandscapeDNDC

In the final contribution to my thesis I evaluated the biogeochemical model LandscapeDNDC with my own measurements of the experimental catchment of the Vollnkirchner Bach catchment in the municipal Hüttenberg (50°29′56″ N, 8°33′2″ E) in Germany (see chapter III).

Houska, T., Kraus, D., Kiese, R. and Breuer, L.: Constraining a complex biogeochemical model for multi-site greenhouse gas emission simulations by model-data fusion, Biogeosciences Discuss, 2017, 1–28, doi:10.5194/bg-2017-96, 2017.

Starting in 2008, the catchment is equipped with sensors, with the goal to establish an interdisciplinary landscape-based teaching facility at the Justus Liebig University Giessen. Since than a number of measurements have been performed with point data of pH, C/N, bulk density, saturated conductivity and porosity and continuous time series of meteorological data (precipitation, relative humidity, air temperature, radiation and wind speed), groundwater table, discharge and in-stream nitrate concentration (Aubert and Breuer, 2016; Lauer et al., 2013; Orlowski et al., 2014). The land use in this catchment is mainly dominated by arable land (35%) and forests (37%). Grassland sites (meadows and wetlands, 11%) are distributed along the streams. Settlements and streets cover the rest of the catchment.

For biogeochemical model initialization, calibration and validation, a number of field measurements are required. Typical data include soil moisture (Figure 5) and GHG emissions (Figure 6) as well as site-specific farm management data on arable land (type of crop, ploughing times and depth, fertilization times and amount) and grassland (grazing and cutting times). I initiated these type of measurements at the beginning of my three years project in November 2013. Together with student assistance, I measured soil GHG emissions (N_2O, CO_2 and CH_4) with non-steady state opaque chambers each covering an area of 0.12 m² soil. In total 40 chambers are setup on three arable, two grassland and three forest transects, each consisting of five measurement points on every measurement day (Figure 6).

Weekly sampling is performed with five replicated chambers per transect, following the cost efficient gas-sample-pooling-technique, developed by the Karlsruhe Institute for Technology (Arias-Navarro et al., 2013). According to this approach, at five time intervals of 10 minutes (t0-t4) 10 mL headspace sample are collected subsequently from any of the five replicated chambers and are pooled into one gas tight glass vial. Samples are automatically analysed via gas chromatography (SRI Instruments equipped with an auto-sampler). The soil moisture data is measured under arable land, grassland and forests land use in 15 min intervals (Figure 5).

Figure 5. Measurement devices in the Vollnkirchener Bach catchment. GHG measurements are performed with dark closed chambers (upper panel) with corresponding soil moisture measurements (5TE FDR-sensors, Decagon) under forest (0.05, 0.15, 0.25 and 0.35 m depth), grassland (0.1 and 0.25, 0.4 and 0.55 m depth) and arable (0.2, 0.4, 0.6 and 0.8m depth) land use.

In this study, my particular interest is to constrain the C and N balance with comprehensive input data for the hydrology, biogeochemical and plant physiology modules of LandscapeDNDC. Based on the previous model structure uncertainty analysis (chapter III), I selected a combination that showed fair results and appropriate model runtime. Accordingly, the modules MeTr^x and wcDNDC, were selected. Based on a GLUE analysis, I accepted only model runs (out of n = 100,000 derived with Latin Hypercube sampling), which are within the best 5% of all simulated RMSEs in terms of the respective variable water filled pore space (WFPS) in different depths on arable land, grassland and forest, as well as yield on arable land. In order to achieve realistic GHG simulations from the biogeochemical module MeTr^x of LandscapeDNDC, I took the posterior parameter boundaries of the

soil moisture and vegetation calibration and ran GLUE (n = 100,000) again. This time, I considered the best 5% of all RMSEs in terms of respective N_2O and CO_2 emissions for each land use (A1-3, G1 and W1-3). Again, only the 5% best parameter sets were accepted per land use. Current model and measurement results for the different criteria used in this study are outlined in the following:

Soil moisture: First model runs show that the simulations are in a reasonable agreement with observation data. Simulated soil moisture, given as WFPS in different soil depths, follows the dynamic and magnitude of the field measurements. Only the short-term dynamic is not captured well. We know from previous work, that including the parameters alpha and n of the Van Genuchten retention curve (Mualem, 1976; Van Genuchten, 1980) during model calibration can improve the results to a certain extend (Houska et al., 2014). The simulation of the forest WFPS, however, has some problems in the magnitude of soil rewetting processes. The remaining errors which most likely result from uncertain rainfall data on all land uses (e.g. beginning of July 2014, where all observations indicate a rising soil moisture and simulations do not react) cannot be fitted with parameterization of soil hydraulic processes only. Such errors remain after calibration. Including rainfall data spatial uncertainty might help to improve results at these points.

CO_2 emissions: The dark ecosystem respiration simulation data shows highest emissions across all three land use types in the summer months. Measurements vary between 0 to 200 on arable land, 0 to 69 on grassland and 0 to 19 kg C ha^{-1} day^{-1} in forest land use. Overall, a good fit can be reported. Remaining errors are due to failing soil respiration process after harvest in LandscapeDNDC and remaining uncertainty, which might be related to the curbed dynamic of the soil moisture simulations. Beside an improved soil moisture simulation process, a misinterpretation of the spatial precipitation input signal can also have potential relevance for improving simulations. Remaining parameter uncertainty is in the range of measured uncertainty, with 10, 8 and 4 kg C ha^{-1} day^{-1} for the different land uses, respectively.

N_2O emissions: Measurements vary between 0 to 0.18 on arable land, -0.002 to 0.014 on grassland and -0.002 to 0.013 kg N ha^{-1} day^{-1} on forest land use. The dynamic of the measured N_2O emissions was reproduced reasonable well by the model on the arable study site. The grassland study site seems to be influenced by N input through groundwater, which was not included in the current LandscapeDNDC model setup. The model also failed to reproduce the measured temporal pattern. A mean annual N gap of 41 kg N ha^{-1} was simulated on this land use. The emission pattern on the forest site differs from the model predictions. Main reason for the mismatch are negative emissions, which cannot be predicted by the model due to a missing model process of microbial N_2O uptake (Figure 7).

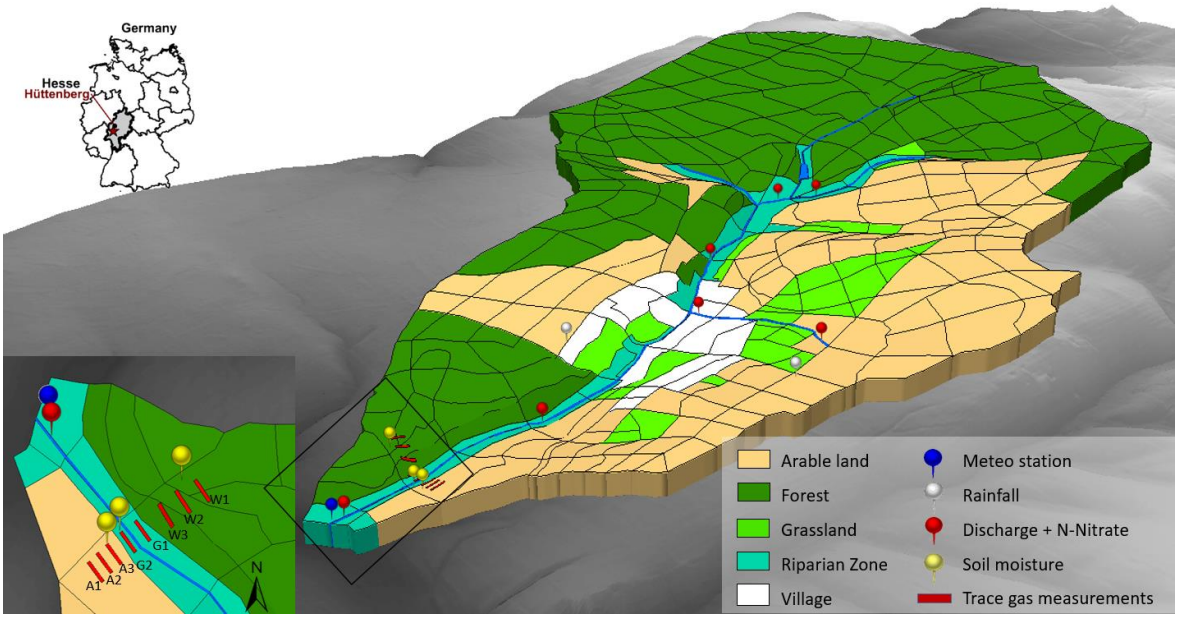

Figure 6. Digital elevation and land use map of the Vollnkirchener Bach catchment-model. The map further shows available sampling sites and measurement data. Continuous high-resolution data measurements for nitrate in stream are available at the outlet and can be used as validation data.

Furthermore, posterior model runs allowed quantifying magnitude and uncertainty of not measured fluxes of the C and N cycle. In general, the investigated forest site is acting as the largest sink for C and N of all studied land uses, with annual sequestration rates of 2.4 t C ha^{-1} and 3.3 kg N ha^{-1}. The extensive grazed grassland is also acting as a sink for C with 1.4 t C ha^{-1} per year, while the N cycle of the grassland model cannot be closed with the given settings. Shrinking N soil pools indicate a missing input, which we assume from shallow groundwater with additional N supply of around 40 kg N ha^{-1} a^{-1}. While the C cycle on the arable land system is closed with low uncertainties, the N cycle is driven by large uncertainties and it remains unclear, if the underlying N pools shrink.

Under the viewpoint of climate smart landscapes, measured data suggests the benefit of forests in a landscape, having the least GHG emissions. Riparian zones can act as sinks of N, but only during the vegetation period and times when roots have access to groundwater. Arable land use produces high amounts of N_2O, but not throughout the year, rather in spring after fertilizer application or during freeze-thaw cycles.

In an overall picture, the model-data fusion approach allowed us to derive missing model processes that would potentially increase model simulation performances if implemented in the respective modules of Landscape DNDC: N_2O uptake processes through microbes; missing NO_3^- (and potentially dissolved organic nitrogen) uptake through shallow groundwater; missing lateral interaction at hillslopes due to 1-dimensional model setup.

While the first point could also be an measurement artefact, is the second point now included in LandscapeDNDC (Liebermann et al., under review). The third point can only be achieved with a spatial application of the LandscapeDNDC-CMF model, i.e. three-dimensional modelling. Such a setup is currently developed, by overlaying soil type, land use and digital elevation maps to create an input map with 352 polygons for the Vollnkirchener Bach catchment (Figure 6). As a meaningful validation of an up-scaled model requires more data, we aim to utilize discharge and nitrate concentration measurements of the outlet of the Vollnkirchener Bach catchment. Further, data on yields and management from farmers working in the catchment will complement the data set. Stream discharge and nitrate concentrations are measured in 15 min resolution with an RBC flume and a UV-Hyperspectral Photometer (ProPS, TriOS, Rastede, Germany), respectively. First results show the general capability of the model to reproduce the catchment discharge and instream nitrate loads. However, the dynamic of the nitrate concentration is not yet accurate in line with observations. We see an overestimation of instream nitrate concentrations during the vegetation period, which is likely due to missing N transformation in the riparian zone. This can be potentially improved, if critical zones are simulated in the coupled mode, i.e. using the more complex Richards equation of CMF. Less critical areas, where capillary rise and groundwater interaction with the plants are not relevant, do not necessarily improve in the coupled mode (Klatt et al., 2017) and could be simulated using the more simple bucket type hydrological approach, resulting in less computational time.

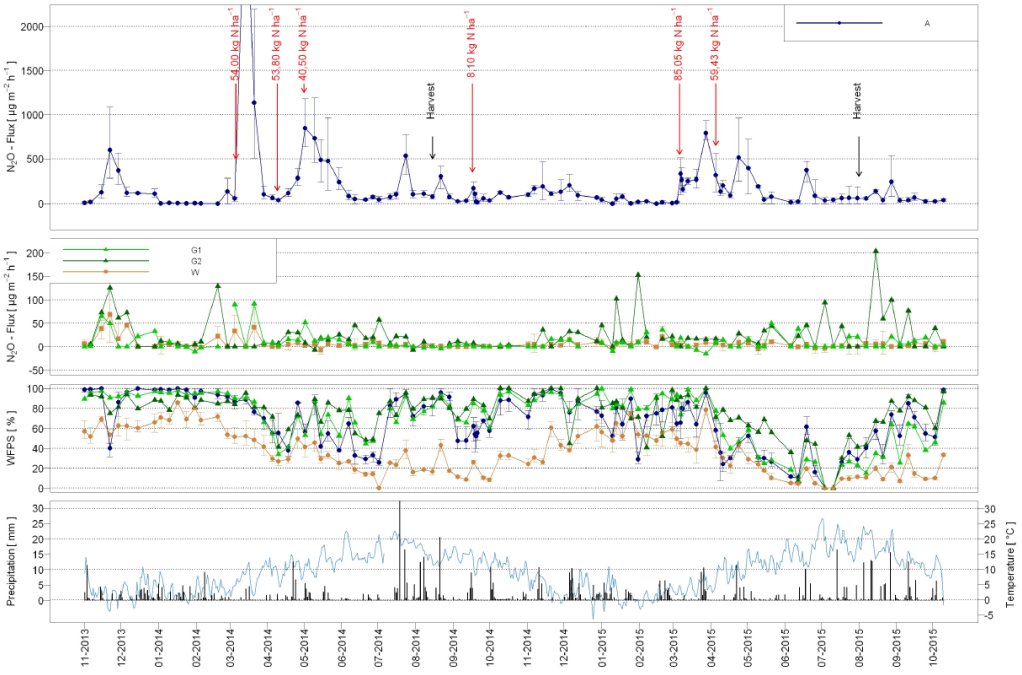

Figure 7. Measured N_2O emissions, precipitation and water filled pore space (WFPS) in the Vollnkirchener Bach catchment with the amount of fertilizer application (red arrows) and harvest dates (black arrows). Variation of the different transects on the land uses A= arable, G = grassland, W = forest, is given as error bars.

Conclusion and outlook

Within my thesis I developed and tested methodologies for assessing the overall model uncertainty of the one-dimensional (only vertical fluxes) biogeochemistry model LandscapeDNDC and of the version, which is coupled to the hydrological model (LandscapeDNDC-CMF). A toolbox to access uncertainty of models was developed and is applied in an increasing number of projects. The uncertainty evaluations of LandscapeDNDC show, that the model can reasonable well, in terms of magnitude and temporal dynamics, represent observed GHG fluxes, biomass production and leaching. However, care needs to be taken even if the framework reproduces measurements well, as it has very limited performance in fitting multiple outputs such as biomass yields, soil GHG fluxes and soil moisture at the same time. My results indicate that the models used in this work still lack on robustness to generate outputs for multiple ecosystem services and I am anticipating that this is the same for many other, complex modeling systems that capture a larger number of environmental processes, fluxes and states. Future work is required to reduce uncertainties and increase model prediction capacity. This requires on the one hand a reduction of model parameters. On the other hand, a change of awareness of site supervisors as data gathering and small-scale changeability of site properties were causing part of the low performance in terms of predicting capability of the Landscape-CMF framework.

Ultimately, the question remains how future model applications can be improved. Based on the results of our model-data fusion approach we were able to show, that the highest uncertainty in the model is due to nitrate leaching and biomass amounts, which could not be sufficiently constrained with measured GHG emissions only. We further need a catchment based hydro-biogeochemical setup, which is covering the relevant spatial scale. As this will result in decreased model runtime performance, we will need to work on advanced model diagnostics techniques to speed up multi-dimensional parameter search. Under the viewpoint of future work, I see the following key research areas:

Improved measurements: Common one-dimensional model setups do not cover key processes of lateral water and N transport (Houska et al., 2017a). Therefore, I recommend to extend the current measurements approaches with the following steps to advance their use for model calibration. Additional data appropriate for model testing would be NO_3^- in soil solution to estimate NO_3^- leaching. Glass suction cups are suited to collect soil water in different depths. I plan an installation in 0.3, 0.5, 1.0, 1.5 and 2.0 m at the arable land study site (A1 and A3) with sampling on a monthly basis. Another sink of C and N is the biomass. A fast and cheap method would be to estimate the biomass dynamics with weekly nondestructive measurements of Photosynthetically Active Radiation

on grass and arable land with absorption sensor (e.g. Delta-T SunScan). Photosynthetically Active Radiation can be used to estimate LAI (Sone et al., 2009). Linear regression can then be applied to convert LAI values to above ground biomass (monthly destructive measurements). The method has been tested e.g. by Goswami et al. (2015) and reported to work well for forests and crops, but has shown limitations in extensive grasslands (Metzger et al., 2017). Samples from aboveground biomass will be measured for C and N content and its ratio for grain, straw and leaves, which are needed for model setup and validation.

Improved model diagnostics: This project showed that multi-objective calibration can be more appropriate than single-objective relationships. However, it does not guard against epistemic errors due to incomplete and/or inexact process knowledge. Explicit knowledge of the various uncertainty sources will provide strategic guidance for investments in data collection and/or model improvement. Thus, I think that communicating the uncertainty of model predictions is a key component of risk-based design and management of landscapes. It enables decision-makers to assess the likelihood that their investments will produce the desired outcome (e.g., reduced nitrate loads and decreased GHG emissions). For the hydrological community, Kavetski et al. (2006a) have developed a way to deal with one further source of uncertainty: the input data. They found the measurement of precipitation within a catchment to be uncertain, as the trajectory of storm cells through a catchment may be different for each storm and may not have its center at the rain gauge where the traditionally rainfall inputs are being measured. This method can significantly improve the rainfall-runoff simulations (Kavetski et al., 2006b). Recently, McInerney et al. (2017) tested eight common residual error schemes, i.e. WLS, log-schemes and Box-Cox, to quantify predictive uncertainty of different hydrological models. They found the choice of residual error model as a significantly impact on predictive performance. I am not aware of any study quantifying predictive uncertainty of hydro-biogeochemical flux simulations so far. I included both ways to quantify input and predictive uncertainty into our modelling scheme during a cooperation visit to Prof. Dmitri Kavetski, University of Adelaide, Australia in autumn 2016, paid by a DFG grants (BR2238-27). I plan to quantify these uncertainties in my calibrated model outputs in order to guide further model development and field measurements.

Reduced nitrogen pollution at landscape scale: Chemical fertilizer application or the recycling of municipal waste on farmland are punctual events that affect soil dynamics and have long-term consequences, which can be predicted by models. However, can models guide to the best ecological and economic management practices? Maximizing yield while minimizing environmental nitrogen losses is in the center of environmental pollution research. The arable study site in the Vollnkirchener

Bach catchment is managed by a cooperative farmer and shows comparatively high annual N_2O emissions of 4.5 kg N ha^{-1}. I plan to explore scenario studies on the effectiveness of different agricultural practices for their potential to reduce environmental N losses. I will focus on soil N_2O emissions and nitrate leaching, while maintaining yields. Scenarios will be first developed and tested using the coupled model LandscapeDNDC-CMF. A subsequent real world optimization approach will be followed in a prognostic way to predict optimal timings and fertilization options (rates and splitting) upon accurate weather forecasts. Potential effects on GHG emissions and nitrate leaching will be monitored.

Overall, I am confident that these improvements will help to make the LandscapeDNDC-CMF model network finally ready to be used in scenario analysis, e.g. to reduce nitrogen pollution through mitigated management on the landscape scale. The tools and approaches generated in this joined project became essential for model evaluation in both working groups involved in this study (Prof. Breuer, JLU Gießen, and Prof. Butterbach-Bahl, KIT, Garmisch-Partenkirchen) and beyond. What has been achieved in terms of assessing the uncertainty of uncoupled and coupled biogeochemical and hydrological models was path breaking and has not been done before. The model development is clearly profiting from the uncertainty assessment as it is used for guiding future work, e.g. a model-data fusion approach comprising the main land uses arable, grassland and forest in a developed landscape (Houska et al., 2017a). Findings presented in this dissertation can now guide the analysis of environmental models. It further allows to investigate long-term measured data of GHG emissions and application of robust hydro-biogeochemical modelling approaches.

Data availability

All measured data is available upon request from the institutes own database
http://fb09-pasig.umwelt.uni-giessen.de:8081

The hydrological model build with the Catchment Modelling Framework (CMF) is free available
http://fb09-pasig.umwelt.uni-giessen.de/cmf

The biogeochemical model framework (LandscapeDNDC) is available upon request
http://svn.imk-ifu.kit.edu

The new developed statistical parameter optimization tool (SPOTPY) is free available from the official Python package repository
https://pypi.python.org/pypi/spotpy

I. SPOTing Model Parameters Using a Ready-Made Python Package

This chapter is published in the journal "PLoS ONE" written by:

Houska, T.[1], Kraft, P.[1], Chamorro-Chavez, A.[1] and Breuer, L.[1,2]: SPOTting Model Parameters Using a Ready-Made Python Package, PLoS ONE, 10(12), e0145180, doi:10.1371/journal.pone.0145180, 2015.

[1] Institute for Landscape Ecology and Resources Management (ILR), Research Centre for BioSystems, Land Use and Nutrition (IFZ), Justus Liebig University, Giessen, Germany
[2] Centre for International Development and Environmental Research (ZEU), Justus Liebig University, Giessen, Germany

Abstract. The choice for specific parameter estimation methods is often more dependent on its availability than its performance. We developed SPOTPY (Statistical Parameter Optimization Tool), an open source python package containing a comprehensive set of methods typically used to calibrate, analyze and optimize parameters for a wide range of ecological models. SPOTPY currently contains eight widely used algorithms, 11 objective functions, and can sample from eight parameter distributions. SPOTPY has a model-independent structure and can be run in parallel from the workstation to large computation clusters using the Message Passing Interface (MPI). We tested SPOTPY in five different case studies to parameterize the Rosenbrock, Griewank and Ackley functions, a one-dimensional physically based soil moisture routine, where we searched for parameters of the van Genuchten-Mualem function and a calibration of a biogeochemistry model with different objective functions. The case studies reveal that the implemented SPOTPY methods can be used for any model with just a minimal amount of code for maximal power of parameter optimization. They further show the benefit of having one package at hand that includes number of well performing parameter search methods, since not every case study can be solved sufficiently with every algorithm or every objective function.

Introduction

Ecological models are often very complex and contain many parameters that need to be optimized prior to model application. Reliable parameter estimation is highly dependent on various criteria, including the selected algorithm, the objective function and the definition of the prior parameter distribution. Difficulties involved in calibrating for example hydrological models have been partly attributed to the lack of robust optimization tools (Duan et al., 1994). Numerous parameterization methods have been developed in the past (e.g. (Bárdossy and Singh, 2008; ter Braak and Vrugt, 2008; Kirkpatrick et al., 1983; McKay et al., 1979; Metropolis et al., 1953)), often published without access to the source code. They are widely accepted to determine the values of non-measureable parameters for a model (Schuëller and Pradlwarter, 2007). Many of the methods have been established as part of the parameterization problem in hydrological modeling as early as in the 1990s (Efstratiadis and Koutsoyiannis, 2010; Matott et al., 2009). The application of these methods has now become more

widespread in other ecological disciplines and therefore, the methods proposed here, are in fact applicable to a large variety of models in ecology and beyond.

The main goal of parameter optimization is to find one or more sets of parameters, which enables a model to simulate an output with a quasi-optimal objective function. There have been extensive discussions about the best way of model parameterization and calibration (Beven and Freer, 2001), including dispute about whether there is one optimal parameter set or whether there are several parameter sets of equal behavior (equifinality, (Beven, 2006)). The same is true for the discussion of the best likelihood function to be used (Smith et al., 2010), how it is determined (Smith et al., 2015) and the parameter distribution from which parameters should be sampled (Haan et al., 1998). Furthermore, improper application of calibration methods can result in misleading parameter estimations (Kavetski et al., 2006). However, nearly no guidance exists which parameter estimation method should be used under specific optimization problems (Wagener and Gupta, 2005). We want to contribute to these open questions by providing a package that allows investigation of various aspects in model calibration, parameterization and uncertainty analyses. The goal is to help users in testing and finding an efficient technique for their specific parameter search problem.

Numerous ad hoc solutions for the combination of a single calibration/uncertainty method and a single model exist. If one is interested in testing different methods, every solution has to be searched, understood and adopted. This is why in recent years packages where published, providing multiple methods for multiple models. Important ones are: Parameter ESTimation and uncertainty analysis (PEST) (Doherty and Johnston, 2003), the Monte Carlo Analysis Toolbox (MCAT), a parameter estimation toolbox (Yang et al., 2008) for the Soil Water Assessment Tool (SWAT), OpenBUGS (Bayesian inference Using Gibbs Sampling) (Lunn et al., 2000), STAN (Hoffman and Gelman, 2011) and PYMC (Patil et al., 2010). However, most of these packages only allow usage of two or three multiple stochastic probabilistic methods. Packages like STAN and PYMC concentrate on Markov Chain Monte Carlo (MCMC) methods. PEST bridges the gap to evolutionary computation methods, a second group of probabilistic global optimization methods (Weise, 2009), like e.g. Shuffled Complex Evolution (SCE-UA) (Duan et al., 1994), but has no possibility to use e.g. the Generalized Likelihood Uncertainty Estimation method (GLUE) (Beven and Binley, 1992; Beven and Freer, 2001), which is widely used to address the equifinality problem. MCAT helps to use the GLUE methodology for models. None of these packages covers the wide range of available parameter search methods. Further, no common criteria exist that place the development of such packages in a formal framework. We therefore define five criteria, inspired by the criteria for modern hydrological models (Buytaert et al., 2008), which we think are important:

1) Broadness: The available parameter estimation methods should cover a broad range of method families, ranging from path-oriented optimizations to global parameter behavioral uncertainty assessments. This is even more important as no single parameter estimation technique is perfect (Jung et al., 2006; Sorooshian et al., 1993) and just very small guidance exists, which parameter estimation approach should be used under specific circumstances (Wagener and Gupta, 2005).

2) Modularity: Parameter estimation packages consist of several modules: the parameter search algorithm, the objective function, a module to save the results of the model runs to disk and the used model. By using a strict modular approach, any given search algorithm can easily be combined with any objective function, giving the user the maximum freedom to adopt a method to a given scientific question.

3) Independency: A model independent package facilitates widespread applications. While a method that is bound to a given model can be used to explore parameter uncertainty, structural model uncertainty remains unquestioned. A model independent method allows the comparison of different model structures using the same parameter space exploration technique and hence the comparison of model structural errors.

4) Scalability: This requirement is an extension of the portability claim (Buytaert et al., 2008). While we agree, that published codes should run both on Windows PC for method testing, as well as on Linux based high performance computing (HPC) systems, scalability goes beyond the portability claim. Scalability means on the one hand, a simple parallelization of the parameter search, where the algorithm allows parallel computation. A package should allow using the parallel power of HPC systems without the need for extensive knowledge of parallel systems. On the other hand, scalability means the possibility to optimize the computational performance of the model. The runtime of models that are fast to evaluate, like e.g. HBV (Bergström et al., 1995) is often dominated by the time needed to load the parameters and input data from disk, and not by the CPU time. Tweaking the model to accept input data through memory can speed up the model evaluation by a magnitude. A scalable package should therefore allow such optimizations and not rely on input file manipulation as an interface between the parameter estimation method and the model alone, as it is the case for most model independent estimation packages.

5) Accessibility: Since a broad, modular package for parameter-estimation carries already a generalized infrastructure for parameter estimation, publishing the package as a free software enables method developers to extend it, without the need to reinvent for example likelihood definitions or parallelization structures. As such, new methods using the existing infrastructure can easily use all

existing methods without further development. However, making the source code available for the public is not sufficient for accessibility. The source must also be modular in its structure and well documented, to simplify the adoption of the underlying infrastructure.

We have developed the parameter-spotting package SPOTPY in agreement with these five criteria. We have implemented and tested a wide range of commonly used algorithms into SPOTPY, to allow a user-friendly access to these powerful techniques, and to give an overview, which algorithms and which objective functions can be useful under specific parameter search problems.

Methods

Concept of SPOTPY

SPOTPY is broad as it comes along with different global optimization approaches. We included the Monte Carlo (MC), Latin Hyper Cube Sampler (LHS) (McKay et al., 1979) and Robust Parameter Estimation (ROPE) (Bárdossy and Singh, 2008) methods that belong to the first group of stochastic probabilistic methods. They are all-around algorithms, applicable for uncertainty and calibration analysis. MC and LHS can furthermore be utilized within the GLUE methodology. Simulated Annealing (SA) is a heuristic subgroup of the stochastic probabilistic methods. We included a version by Kirckpatrick et al. (Kirkpatrick et al., 1983). The Maximum Likelihood Estimation method (MLE) belongs to the subgroup of hill climbing algorithms and is suited for monotonic response surfaces. Markov Chain Monte Carlo (MCMC) methods, a subgroup of the probabilistic methods, support the ability to jump away from local minima. We implemented the standard Metropolis MCMC sampler (Metropolis et al., 1953). To cover the second group of probabilistic methods (evolutionary algorithms) we included the evolution strategy of SCE-UA. It is suited to calibrate models with high parameter space. Furthermore, the Differential Evolution Markov Chain (DE-MC_Z) was included to provide a Bayesian solution suited for optimization problems in high parameter space.

SPOTPY is modular since prior parameter distributions, model inputs, evaluation data and objective functions can be selected and combined by the user. The user-defined combination of the inputs can be run with the parameter search algorithms and results are saved either on the working storage or in a csv file. The database structure enables the analyses of the results in SPOTPY with pre-build plotting functions and statistical analyses like Gelman-Rubin diagnostic (Gelman and Rubin, 1992)

or the Geweke test (Geweke, 1992). The database can also be used for any other external statistical software or computer language.

SPOTPY is independent as the model is wrapped in a “black box”. One parameter set is defined as input; the model results are defined as output. Both deterministic and stochastic models can be analyzed.

SPOTPY scalability is realized by using the Python programming language, since it has an increasing support from the scientific community and is a recommended programming language for scientific research (Perkel, 2015). Pure Python code can run on every operating system without any complicated building mechanism. Parallel computing on HPC systems is supported by using a Message Passing Interface (MPI) code. Five of the eight implemented algorithms are suitable to for parallel computing (MC, LHS, SCE-UA, DE-MC$_Z$, ROPE). The MPI code depends on the open source python package mpi4py (Dalcín et al., 2008). A sequential run does not have any dependencies to non-standard python libraries. SPOTPY is accessible as open-source on the Python package index PyPI and comes along with tutorials to allow a user-friendly start without the need of a graphical user interface and the benefit that everyone can use the most recent version of the code (Moeck et al., 2015). The code follows object-orientated style, where it supports modularity and is conform to the Open Source Definition (Anon, 2015).

Structure of SPOTPY

The design of SPOTPY brings different parameter estimation approaches within one set-up to allow users testing a variety of different combinations and methods. Figure I.1 shows the main processes of this package, consisting of six consecutive steps when applying SPOTPY.

The different steps included are the following:

Step 1) Parameter distribution: Let $\theta = \{p_1, p_2, \ldots p_n\}$ be the initial input set of parameters of a (ecological-) model *M*. The $\{p_i\}_{i=1}^{n}$ random variables are selected from a joint probability prior distribution. This can be any user-defined distribution. We have pre-built the distributions Uniform, (log-) Normal, Chi-square, Exponential, Gamma, Gamma, Wald and Weilbull with NumPy (Oliphant, 2006). Each parameter p_i is marked with a user defined name, step size and optimal guess (initial parameter set), which are used as prior information by the algorithms and the database. The parameter name is used by the database, while the step size is an information needed for MCMC, MLE and SA to jump to the next point of the prior distribution. The optimal guess is the start point for all algorithms. The better this value is chosen, the faster convergence can be achieved.

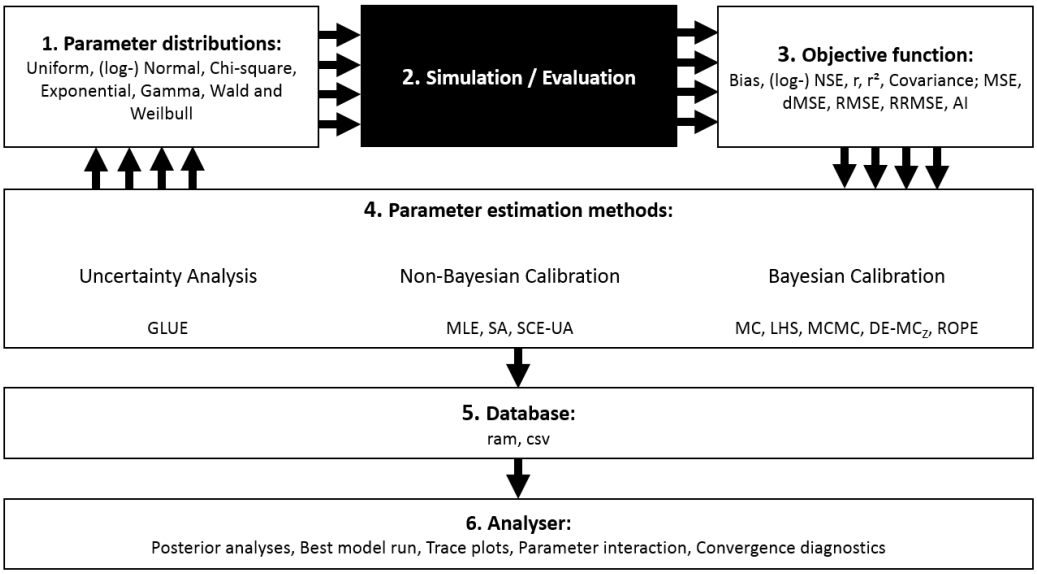

Figure I.1. Flow diagram of the main processes captured with SPOTPY. Multiple cycle black arrows indicate the possibility of parallelization of the iterating algorithms. The black box returning the simulation and evaluation data can be filled with any model.

Step 2) Simulation and Evaluation**:** The output of *M* given a parameter set θ_i is defined as simulation *S*. The observed data *X* is characterized as evaluation. The simulation function is designed to call a model, returning a list of simulated values. The observation data is loaded in the evaluation function. One can also analyze a model with SPOTPY, which is only returning an objective function. Both functions and the following objective function offer the user flexibility to analyze almost every model with SPOTPY.

Step 3) Objective function**:** The objective function (also known as cost-function or goodness-of-fit-measure) quantifies how well the simulated data fits the evaluation data. Various objective functions are available (e.g. (Nash and Sutcliffe, 1970; Willmott, 1981)) and have been proposed to account different sorts of errors in the simulation (Legates and McCabe, 1999; Li et al., 2010). A guidance, which objective function to take under specific circumstances, is given by (Moriasi et al., 2007). Hence, SPOTPY comes along with a wide set of objective functions, from which the user can select one or more for a specific issue (Bias; Nash-Sutcliff efficiency; logarithmized Nash-Sutcliff efficiency; Correlation Coefficient; Coefficient of Determination; Covariance; Decomposed, Relative and Root Mean Squared Error; Mean Absolute Error; Agreement Index). The user has the option to combine different objective functions as only one function can be inaccurate (Vrugt et al., 2003). A detailed description of the objective functions implemented in SPOTPY can be found for example in (Wallach, 2006).

Step 4) Parameter estimation methods: The algorithms included in SPOTPY cover widely used parameter estimation methods from different approaches in recent publications. They can be connected with setup files containing the above-mentioned information about parameter distribution, simulation- and evaluation data as well as the objective functions. The simplest automatic parameter estimation method included is the MC method. It is used to sample random parameter values from a prior distribution. The structural LHS algorithm subdivides the distribution of each parameter into *m* equally probable non-overlapping intervals and creates a matrix by sampling from all created intervals. The algorithm has shown good projection properties (Hossain et al., 2006; Murphy et al., 2006; Over et al., 2015; Windhorst et al., 2014). MC and LHS can form the basis for the GLUE method (Beven and Binley, 1992; Beven and Freer, 2001), to get information about the posterior distribution of input parameters. GLUE has been widely applied in hydrology, but also in many other ecological disciplines, such as biogeochemistry or crop growth modeling (Houska et al., 2014; Ortiz et al., 2011; Shafii et al., 2015). If one is just interested in a fast calibration of a simple model (with nearly monotonically response function), the MLE is an efficient choice. To test whether the MLE algorithm is applicable for calibrating the desired model, it is recommend to test the model with MC first (Kitanidis and Lane, 1985). MLE maximizes the likelihood during the sampling, by adapting the parameter only in directions with an increasing likelihood. The famous Metropolis MCMC method can also deal with non-monotonically response functions. Nevertheless, it works similar as MLE. After each sampling, the likelihood is compared with last one. If the likelihood is better, the sampler jumps to the new sampled point. If not, it samples from the old position. Depending on a Metropolis decision, the sampler can also accept worse likelihoods (in order to avoid trapping at local optima). The MCMC algorithm can find a (quasi-) global optimum, but with a still remaining risk to stuck in local minima. The risk can be reduced by starting several chains/complexes that evolve individually in the parameter space. This technique is used in the global optimization strategy SCE-UA (Duan et al., 1994). Each complex evolves independently to optimize the parameter. The population is periodically shuffled and new complexes are created with information from the previous complex. SCE-UA has found to be very robust in finding the global optimum of hydrological models and is one the most widely used algorithm in hydrological applications today (Over et al., 2015). Another robust method is SA. Thyer et al. (Thyer et al., 1999) reported SA to be not as robust as the SCE-UA algorithm, but SA can be very efficient, when it is adopted to a optimization problem. After each step, a better objective function results in a new position. A worse objective function can be accepted with a Boltzman decision. If the new point is not accepted, the sampler jumps to a new parameter value. A variable controls a decreasing possibility to accept worse objective functions with increasing iterations. Thus, the risk to jump away from a global optimum is reduced. One of the most recent

algorithms we present here is the DE-MC$_Z$. It requires a minimal number of three chains that learn from each other during the sampling. It has the same Metropolis decision as the MCMC algorithm and has found to be quite efficient compared with other MCMC techniques (Smith and Marshall, 2008). Like SCE-UA and SA, DE-MC$_Z$ does not require any prior distribution information. Another non-Bayesian approach is to determine parameter uncertainty estimations with the concept of data depth. This has the benefit, that the resulting parameter sets have proven to be more likely giving good results when space or time period of the model changes, e.g. for validation (Kraußе and Cullmann, 2012). This approach is realized in the ROPE algorithm.

Step 5) Database**:** The database can store results from every parameter estimation method. Either in the working storage, which is fast, or in a csv file, which is comfortable. Saved information for every iteration are the objective function (-s), every parameter setting, optional the simulation results and the chain number (for algorithms with multiple threads like SCE-UA and DE-MC$_Z$). The database can be analysed in any statistical software, programming language or the SPOTPY extension Analyser.

Step 6) Analyser**:** The Analyser module is an optional, but very powerful extension, which can read the SPOTPY database. Prebuild plots are provided for objective function and parameter traces, parameter interactions and best model runs. Posterior parameter sets can be selected and basic statistical analysis of the samples can be performed with Gelman-Rubin diagnostic (Gelman and Rubin, 1992) or the Geweke test (Geweke, 1992).

To install SPOTPY, one just has to type *pip install spotpy* into the OS console. After that, SPOTPY can be used from any Python console:

```
import spotpy                                          #Import the package
from spotpy_setup_rosenbrock import spotpy_setup #Import an example setup
sampler = spotpy.algorithms.sceua(model_setup()) #Initialize an algorithm
sampler.sample(10000)                                  #Run the model n times
results = sampler.getdata()                            #Load the results
from spotpy import analyser                            #Import opt. extension
spotpy.analyser.plot_parametertrace(results)      #Plot the results
```

Set up of algorithms

The setting of the algorithms for the following case studies are depicted in Table I.1. Two things are changed during the case studies: 1) The number of repetitions. 2) For efficiency reasons the set-up of the algorithms was slightly changed when, sampling from the Ackley function in the third case study: SA with Tini=30, Ntemp=30, SCE-UA with ngs=2 and DE-MC$_Z$ with nChains=dim.

All settings of the algorithms should be adjusted, when dealing with other optimization problems.

Table I.1. Settings of the algorithms used in the case studies.

Algorithm	Setting			Source
	Description	**Abbreviation**	**Value**	
MC	Normal random sampling			
LHS	Normal sampling along the HyperCube matrix			[3]
MLE	Percentage of repetitions dedicated as initial samples	burn-in	10%	
MCMC	Percentage of repetitions dedicated as initial samples	burn-in	10%	[4]
SCE-UA	Number of parameters	dim		[1]
	Number of complexes	ngs	2(dim)	
	Maximum number of evolution loops before convergence	kstop	50	
	The percentage change allowed in kstop loops before convergence	pcento	10^{-5}	
	Convergence criterion	peps	10^{-4}	
SA	Starting temperature	Tini	10	[6]
	Number of trials per temperature	Ntemp	10	
	Temperature reduction	alpha	0.99	
DE-MC$_Z$	Number of different chains to employ	nChains	2(dim)	[2]
	Number of pairs of chains to base movements	DEpairs	2	
	Interval to save status	thin	1	
	Factor to jitter the chains	eps	0.04	
	Convergence criterion		0.9	
	Automatic adaption		True	
ROPE	Number of optimization cycles	subsets	5	[5]
	Acceptance ratio	percentage	0.05	

For further detailed description of the SPOTPY package and the presented case studies see the download page (https://pypi.python.org/pypi/spotpy/ and the online documentation http://www.uni-giessen.de/cms/faculties/f09/institutes/ilr/hydro/download/spotpy).

Case studies

We show five different case studies to depict the capability of the different algorithms integrated in SPOTPY under different parameter optimization problems. Three of these case studies cover classical numerical optimization problems with a known posterior target distribution, one a hydrological model simulating real-world measured soil moisture values and one a biogeochemistry model where we tested the influence of different objective functions.

Rosenbrock function

The Rosenbrock function (Rosenbrock, 1960) is often used to test and compare the performance of optimization methods (Goodman and Weare, 2010; Matott et al., 2013; Santos et al., 2000; Wang et al., 2014). It can be described as a flat parabolic valley (Figure I.2) and is defined by

$$f_{Rosen}(x, y) = (1 - x)^2 + 100(y - x^2)^2, \quad (1)$$

where we set the parameter space of the control variables to $x \in [-10,10]$ and $y \in [-10,10]$.

The global minimum is located at $(x_{opt}, y_{opt}) = (1,1)$. At this point the function value is $f_{Rosen}(x, y) = 0$. Due to its shape, it is an easy playground for optimization algorithms to find the flat valley, but it is hard to find the deepest point.

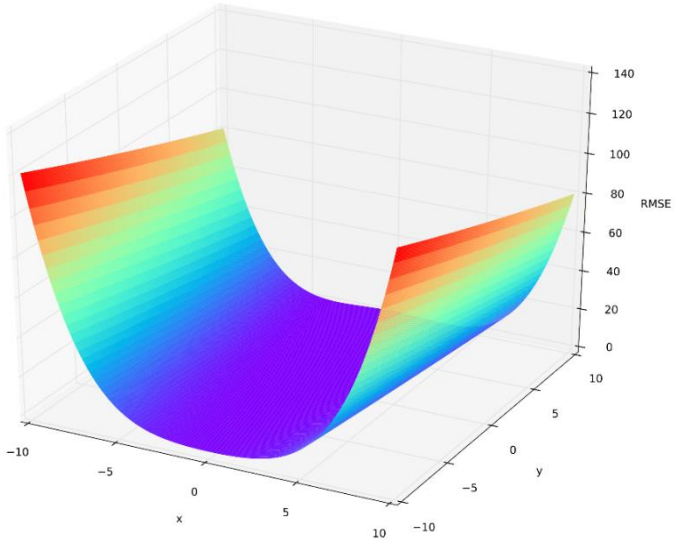

Figure I.2. Three-dimensional surface plot of the Rosenbrock function. Colors from red (bad) to violet (optimal) represent the corresponding objective function (RMSE) for a parameter setting of x and y.

Trace plots were created after sampling n=5,000 times from parameter space of the Rosenbrock function. Figure I.3 depicts the behavior of the algorithms. MC and LHS sample from the complete parameter distribution over the whole time. These algorithms find a few parameter distributions around the global optimum, which are masked by the overall large spread of selected parameter sets. All other algorithms show improved performances with increasing iterations. After 500 runs of burn-in, the MLE algorithm is very fast in finding the region around the global optimum. The MCMC works similar to the MLE, but with the possibility to jump away from the optimum. The algorithm finds the global optimum after 800 iterations and remains with a relative high uncertainty of x=4 and y=4. SA is fast in finding the valley and returns samples with a smaller uncertainty than MCMC. SCE-UA and DE-MC$_Z$ sample in the first iterations over the whole range and converge at the global optimum after 800 and 1,000 iterations, respectively. SCE-UA stops after finding the exact global optimum. DE-MC$_Z$ continues to produce parameter combinations close to the optimum with $x \in [-0.5,0.5]$ and $y \in [-0.5,1]$. ROPE converges systematic closer to the optimum. The *y* variable range is reduced rather quickly to only positive values. For the *x* variable range the convergence works overall better. Overall, it turns out that MLE, MCMC, SCE-UA and DE-MC$_Z$ are the most suited algorithms in finding the global optimum of the Rosenbrock function.

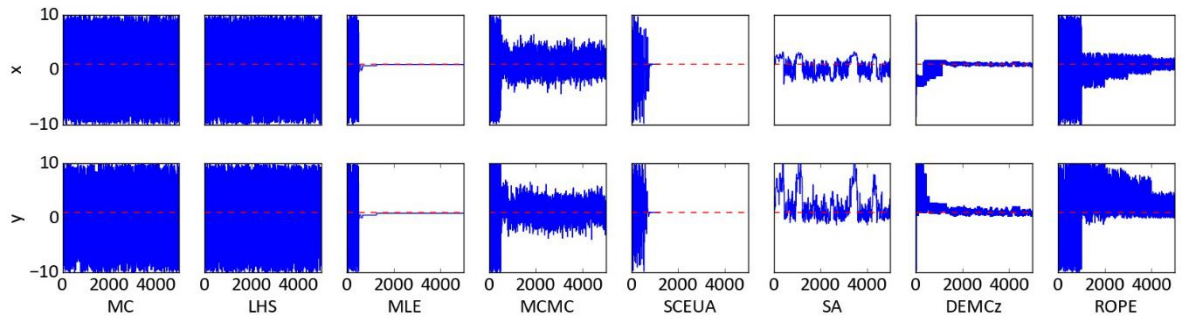

Figure I.3. Trace plot of the two dimensional Rosenbrock function. The trace is shown as a blue line and the global optimum of the function as a broken red line. The x-axes show the number of iterations, while the y-axes show the value of the parameters x and y from -10 to 10.

Griewank function

The two dimensional Griewank function (Griewank, 1981) is defined as

$$f_{Griewank}(x,y) = \frac{x^2+y^2}{4000} - \cos\left(\frac{x}{\sqrt{2}}\right)\cos\left(\frac{y}{\sqrt{3}}\right)+1, \quad (2)$$

where we selected the parameter space for $x \in [-50,50]$ and $y \in [-50,50]$. One of the characteristics of the function is that it has many regularly distributed local minima (Figure I.4), which makes it challenging to find the global optimum located at $(x_{opt}, y_{opt}) = (0,0)$. The demanding function has been used for algorithm performance testing by others (Alfi, 2011; Harp and Vesselinov, 2012; Storn and Price, 1997). The surface of this function allows the investigation the algorithm performance under equifinality.

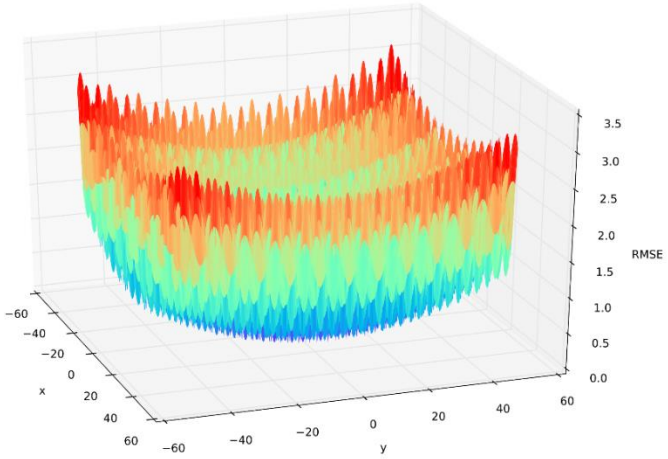

Figure I.4. Three-dimensional surface plot of the Griewank function. Colors from red (bad) to violet (optimal) represent the corresponding objective function (RMSE) for a parameter setting of x and y.

The different algorithms were applied to the Griewank function (n=5,000 iterations). The parameter interactions are shown as combined dotty plots (Figure I.5). We added a surface plot of the Griewank function to show the locations of the various local minima. We conducted the GLUE methodology to MC and LHS by selecting the 10% best runs. One can see samples for MC and LHS on almost every local minima and the global optimum. The random walk of the MLE jumps between three local minima after the burn in, without finding the global optimum. The MCMC algorithm reaches several local minima in intermediate steps and found the global minimum. Nevertheless, the samples orientate not on the local minima and form clouds around the optimum.

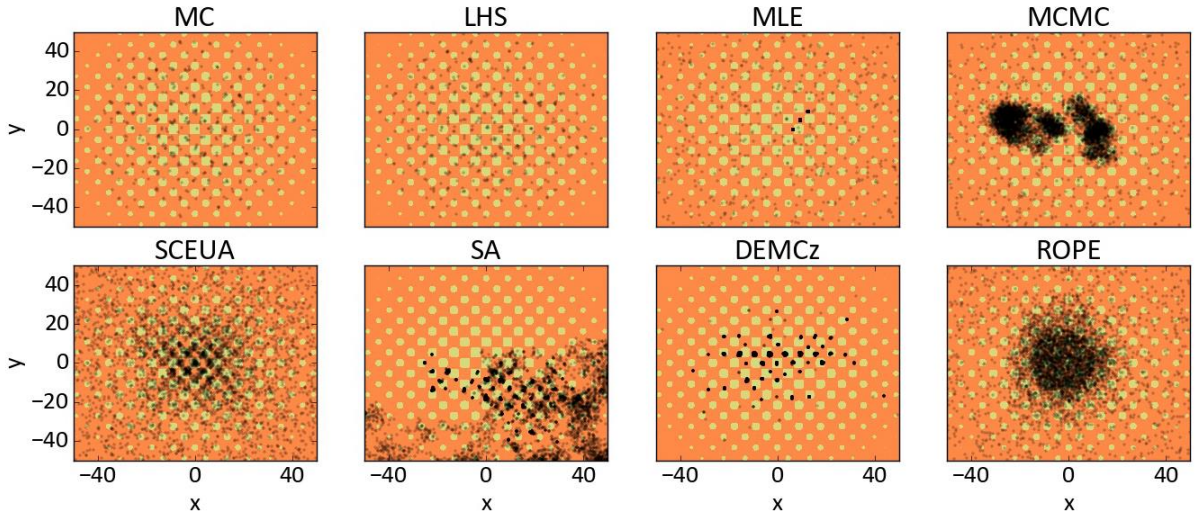

Figure I.5. Surface plot of the Griewank function. Background colours showed from orange (bad response) to yellow (optimal response). Black dots show the sampled 5,000 parameter combinations. The x-axis shows the range of parameter x and the y-axis of parameter y.

The SCE-UA samples parameter combinations from the whole range and reduces the range more and more to the global optimum. It stops the search after 4,000 iterations; nevertheless, the remaining parameter uncertainty is still high. SA did not find the optimal value and samples only negative values for the parameter y. DE-MC_Z found many local minima and the global optimum, which is representing the hilly response surface very good. ROPE reduced the investigated parameter range gradually centered to the optimal point.

Ackley function

The Ackley function is defined as

$$f_{Ackley}(\mathbf{x}) = -20\exp(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d}x_i^2}) - \exp(\frac{1}{d}\sum_{i=1}^{d}\cos(2\tau\tau x_i)) + 20 + \exp(1), \quad (3)$$

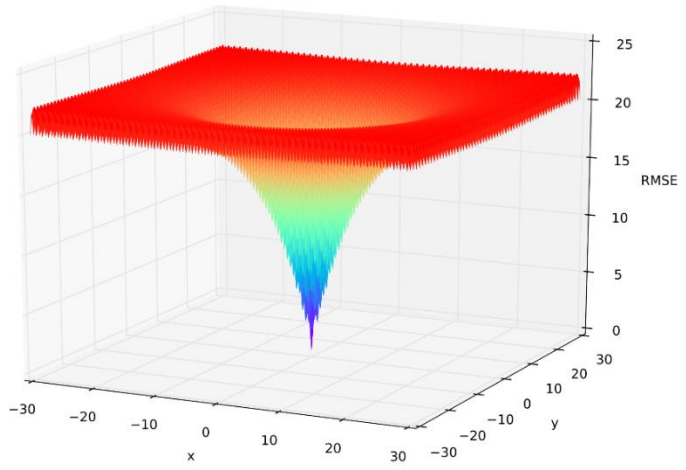

Figure I.6. Three-dimensional surface plot of the Ackley function. Colors from red (bad) to violet (optimal) represent the corresponding objective function (RMSE) for a parameter setting of x and y.

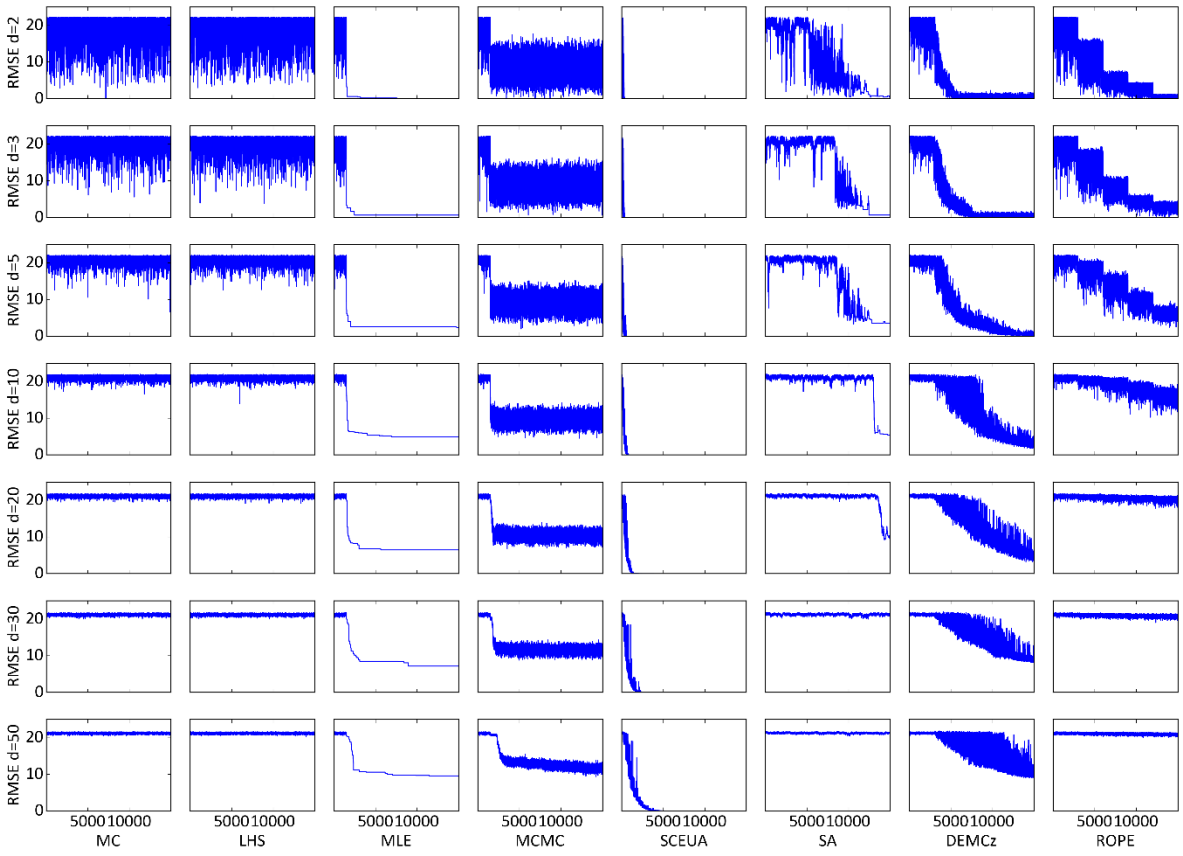

Figure I.7. Objective function traces of the Ackley function. Setup with 2, 3, 5, 10, 20, 30 and 50 domains from the vector x of the Ackley function. All algorithms sampled 15,000 parameter combinations. The shown objective function on the y-axis is the root mean squared error (RMSE). The x-axis shows the number of iterations.

where $\mathbf{x} = (x_i,..,x_d)$ and the domain is defined as $x_i \in [-32.768,32.768]$ (Ackley, 1987). The function has many regularly distributed local minima in the outer region, and a large funnel as the global optimum in the center located at $f_{Ackley}(0,...,0) = 0$ (Figure I.6). The function is widely used for algorithm testing (Potter and Jong, 1994; Stacey et al., 2003; Zhu and Kwong, 2010). We used setups with 2, 3, 5, 10, 20, 30 and 50 domains to investigate the algorithms behavior when dealing with an increasing number of parameters, while finding a very small global optimum.

We used n=15,000 iterations for every setup testing the algorithm's performance (Figure I.7). All algorithms perform worse with increasing dimensions. MC and LHS struggle even with two domains to find the exact optimum. With five domains, MLE, MCMC, SA and ROPE get close to the global optimum but do not find the exact position. With 10 domains DE-MC$_Z$ does not reach the exact global optimum during the 15,000 iterations, but got close with a remaining RMSE of 2-5. With 20 and 30 domains MLE, MCMC and DE-MC$_Z$ still give reasonable results, and can gather information during the iterations to get close to the optimum. Only SCE-UA is able to find the global optimum of the Ackley function with 50 domains during the given number of iterations.

Catchment Modelling Framework

We used the Catchment Modelling Framework (CMF) developed by (Kraft et al., 2011) to investigate the performance of the algorithms when dealing with a real measured world optimization problem. CMF is a toolbox to build water transport models from a set of pre-built process descriptions. The toolbox has been used before to model different catchments in one and two dimensions (Haas et al., 2013; Houska et al., 2014; Kraft et al., 2012; Windhorst et al., 2014) and enables the test of hypotheses in hydrology (Clark et al., 2015). In the application presented here, CMF is set up to simulate soil moisture in a one-dimensional soil column. Evapotranspiration is predicted by the Shuttleworth-Wallace method and soil water fluxes are modeled with the Richards equation. We searched for parameter sets to describe the shape of the water retention curve according to van Genuchten-Mualem (Van Genuchten, 1980) with four parameters: alpha, porosity, n and k_{sat}. The prior parameter distributions are based on results from (Houska et al., 2014), where soil moisture was simulated with CMF for an agricultural site in Muencheberg. We used data from a Free Air Carbon dioxide Enrichment (FACE) grassland study site A1 in Linden, Germany (Jäger et al., 2003). The soil is classified as a Fluvic Gleysol. Meteorological data was used for the weather simulation and groundwater table data for the groundwater influence on this site. For the model evaluation, we utilized daily measured soil moisture data from the topsoil layer (0-0.1 m). The simulation time was

from 01/06/1998 to 01/01/1999 as burn-in and simulation results until 01/01/2000 were used for evaluation.

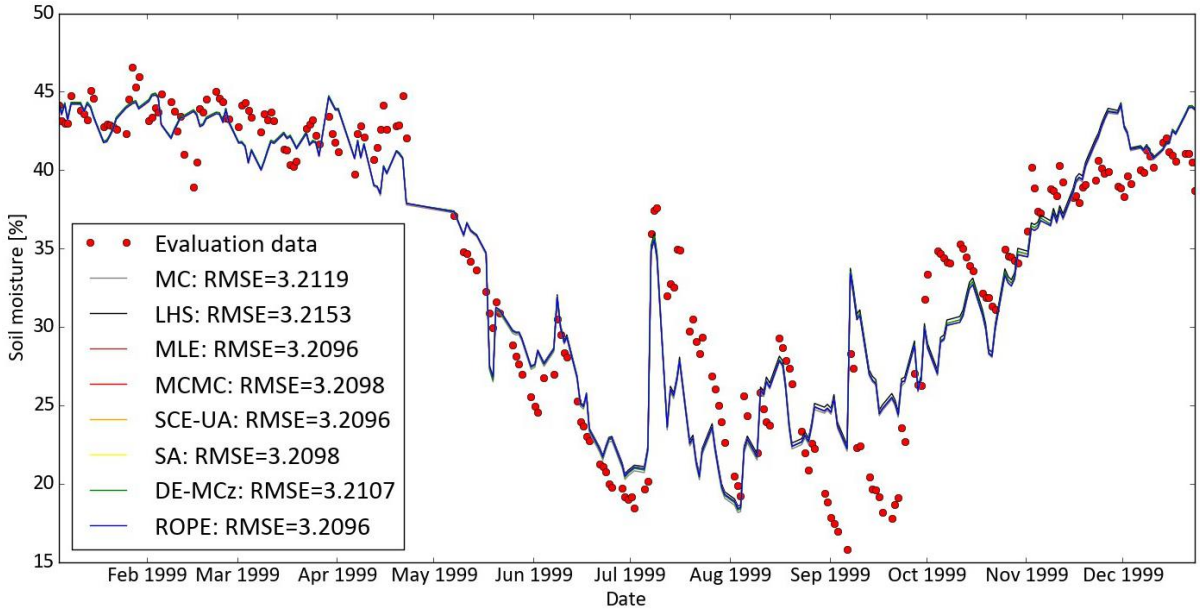

Figure I.8. Best CMF runs for simulating soil moisture. Found with 10,000 iterations of the different algorithms realized with SPOTPY. The resulting different curves are very similar and overlap most of the time.

We started 10,000 iterations with a MPI structure. Twenty parallel threads on a HPC were used, resulting in a nearly linear speed up. The minimal RMSE was used to evaluate model performance. The best model runs of CMF found with the different algorithms are shown in Figure I.8. All algorithms performed almost equally well. The ROPE, SCE-UA and MLE found the best parameter sets for predicting soil moisture with an RMSE as low as 3.2096. All other algorithms performed only slightly worse with RMSE between 3.2098 and 3.2153. Overall, the model simulations follow the main trend of the observations, especially during the first seven months when soil moisture decreased from 45 to 20%. The following flashy soil moisture curve is indicating that the model has deficiencies in simulating rapid changes in soil moisture of the uppermost soil layer, at least with the given forcing precipitation data and available information on soil parameters. This is a problem, which cannot be solved with parameter calibration and needs further investigation, e.g. by improving the model structure, adding more prior information into the process based model, or by testing other models.

Figure I.9 shows the parameter distribution of the best performing parameter sets as well as the prior and posterior distribution (derived by selecting the best 50% of the sampling). The calibration algorithms MLE and SCE-UA resulted in a small posterior distribution. MCMC and DE-MC_Z reduced the parameter uncertainty of the posterior distribution by over 90% for parameter n and by 20% for parameter k_{sat}. The other algorithms failed in reducing the parameter ranges. The optimal

parameter setting for k_{sat} was found on a wide range from 0.8 (MLE) to 1.9 (MC) m day^{-1} and not in the center of the posterior distributions. Optimal settings for the parameter porosity were found in the upper range of the prior distribution, with small posterior distributions. The optimal parameter settings found for alpha, porosity and *n* are close to the center of the posterior distribution.

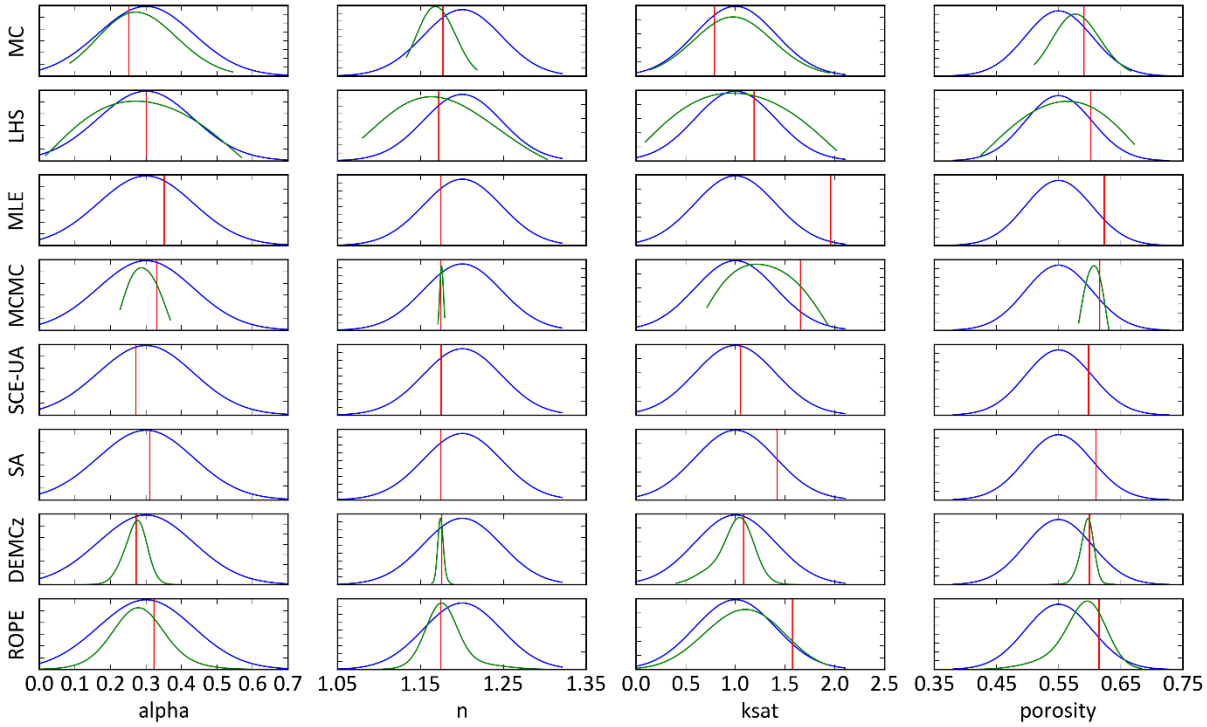

Figure I.9. Prior distribution (blue line) of input parameters of CMF. Posterior distribution (green line) as the best 10% of the samples, plotted only for the Bayesian approaches. The optimal parameter setting is marked with a vertical red line.

We do not know the true optimal parameter set of our hydrological model, or whether it exists at all. The optimal parameter sets we found differ from each other, indicating a high equifinality of the model. The optimal parameter settings for porosity were found in a small range from 0.6 to 0.63 for all algorithms. This values are in line with measured porosity of 0.60 to 0.65 (Kammann et al., 2008). The tested algorithms resulted all in similar best fits, with an RMSE=3.2 Vol. % soil moisture. A direct comparison to other models is not possible, as this is the first study modelling soil moisture on the Linden FACE site. Nevertheless, results are not as good as others, e.g. (Scott et al., 2000) who used SCE-UA and found after 6,000 HYDRUS simulations remaining errors of RMSE=0.03 Vol. % soil moisture on a different site. However, we attribute our relatively high remaining error to model deficiencies in capturing all natural effects, which might be a changing k_{sat} in the upper most soil layer after heavy rainfall on this site (Kammann et al., 2001).

LandscapeDNDC

We used LandscapeDNDC (LDNDC) developed by Haas et al. (2013) to investigate the influence of the chosen objective function on the best selected model run. LDNDC is a biogeochemistry model to simulate greenhouse gas emissions and nutrient turn over processes. We used the model to simulate CO_2 emissions from the soil of the Linden FACE site. The emissions were measured with the closed chamber method (Kammann et al., 2008). We setup the model with a warm-up period of one year and simulated the time from 01/01/1999 to 13/06/2006. Thirty parameters were sampled in a LHS with 50,000 runs. We selected four different widely used objective functions from SPOTPY to quantify the fit of the resulting simulations to the observations (Figure I.10).

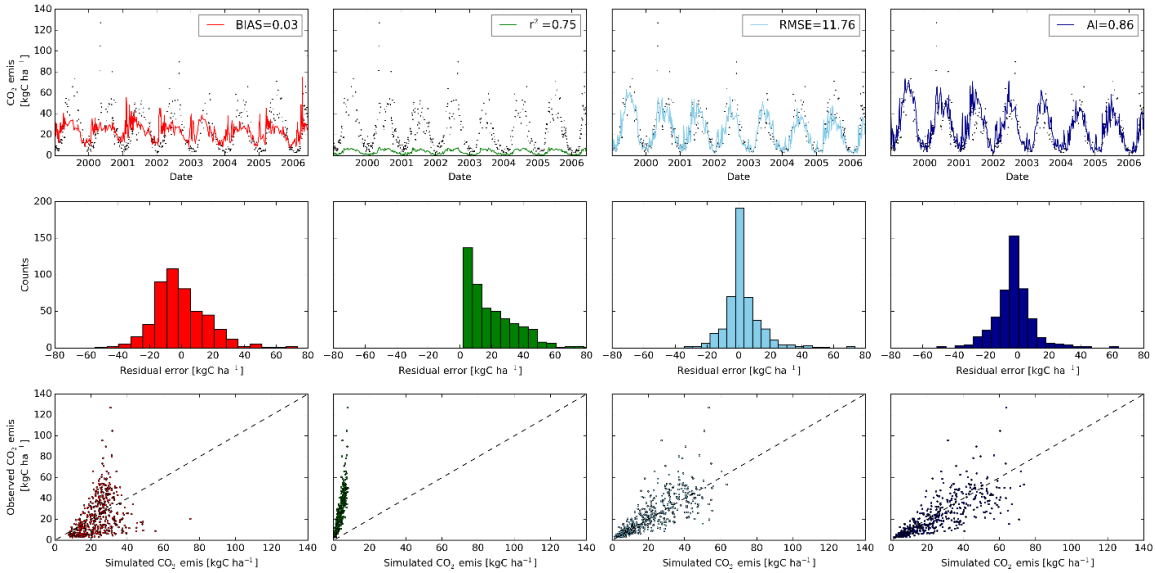

Figure I.10. Comparison of measured and observed CO2 emission simulated with LDNDC (top panels). Best model runs were derived with four different objective functions using a Latin Hypercube sampling approach (n=50,000 model urns). The objective function BIAS is shown in red, r^2 in green, RMSE in light blue and AI in dark blue. Observed values are shown as black dots. Middle panels depict classified residual error counts of simulated CO_2 emissions for each model. The dashed black lines in the correlation plots of observed versus simulated CO_2 emissions (bottom panels) show the theoretical optimal fit.

The selected objective functions were the BIAS (ranging from -∞ to +∞, with 0 indicating an unbiased simulation), r^2 (ranging from 0 total disagreement, to 1 perfect regression), RMSE (ranging from -∞ total disagreement to 0 perfect fit) and the Wilmott Agreement Index (AI, ranging from 0 total disagreement to 1 perfect fit). The best BIAS found has a value of 0.03, which is close to its optimum of zero. However, soil emissions are overestimated in winter with 20 kg C ha^{-1} and underestimated in the summer months with 20 kg C ha^{-1}. Looking at the distribution of the residuals, over- and underestimations are nearly Gaussian, resulting in a mean error near zero over the whole simulation period. The simulation with the best r^2 has a relative high value of 0.75, but the simulations

substantially underestimate the emissions from the soil during the whole model run. Nevertheless, the simulations follow the seasonal trend well, reflecting a reasonable timing of the model (Figure I.10). To improve the fit of absolute emissions with the model, RMSE and AI are good options in SPOTPY. The distribution of the residual errors RMSE are narrower than the ones for AI, which indicates that the observations are better represented by the RMSE optimized model. In contrast, AI optimized simulations are superior in matching the absolute peaks of observed emissions.

Discussion

All algorithms work well in SPOTPY, which was shown by the different case studies. Our intention was not to accept or reject algorithms but rather show their functionality within SPOTPY. Our results show reasonable effects, which have been reported in other algorithm comparison papers. The Rosenbrock case study showed us well performing algorithms when searching for a single optimal parameter set, like MLE and SCE-UA. Vrugt et al. (Vrugt et al., 2003) tested SCEM-UA (similar to SCE-UA) and MCMC on the Rosenbrock function, and reported that the first algorithm was faster in convergence. We found SA struggling in finding the optimum of the Rosenbrock, an observation also reported by Wang et al. (Wang et al., 2014). When dealing with many local minima like it is true for the Griewank function, we got good results, when we conducted MC and LHS with the GLUE concept. They represent best the surface of the function. SCE-UA needed 4,000 iterations to stop the parameter search on the function, Jung et al. (Jung et al., 2006) found the optimum during 40,000 iterations. This difference in efficiency is most likely due to the setting of the algorithm. With an increasing amount of parameters on the Ackley function, we have seen good results for MLE, MCMC and DE-MC_Z and very good results for SCE-UA. Karaboga et al. (Karaboga and Basturk, 2007) tested the swarm intelligence algorithm ABC on the Ackley function with 30 domains. They found after the optimum after 1,000 iterations, which is even better than the best performing algorithm of SPOTPY (SCE-UA). This algorithm could be a nice extension for the SPOTPY package. Behrangi et al. (Behrangi et al., 2008) used SCE-UA in a similar set up and found the optimum of a 30 dimensional Ackley function after around 4,000 iterations, exactly as we found it. Genetic algorithms give poor results on the Ackley function with 30 domains (Karaboga and Basturk, 2007). Madsen et al. (Madsen et al., 2002) calibrated a hydrological model with SCE-UA and SA, showing that the first one worked better – similar to our case studies. Huang et al. (Huang and Liang, 2006) recommend MCMC to deal with many parameters. Our findings on the Ackley function show that evolution algorithms are even better suited for higher dimensional search problems. Ter Braak and Vrugt (ter Braak and Vrugt, 2008) showed that the evolution algorithm DE-MC_Z can be 5-26 times more efficient than MCMC. Gong (Gong, 2006) come to the same conclusion when testing the

evolution algorithm SCEM-UA against the stochastic algorithm MLE. Good results were reported when using MC on a hydrological model with small parameter space (Huang and Liang, 2006). We found that the rather simple MC and LHS often performed worse when searching the exact global optimum, but give reliable results under equifinality, like it is the case for our hydrological model build with CMF. We recommend using these simple search algorithms with the GLUE concept.

The LDNDC case study revealed that conclusions based on the model performance can be flawed when it is analysed with a not well-suited objective function. For example, the BIAS can reduce the overall model error, but it does not guarantee that the model fits the temporal variations of the observed data. The r^2 is suited to find good parameter sets to predict timing of the system, but this objective function does not take the absolute values into account. RMSE and AI are well suited to find model realizations fitting the absolute values of the observed data. Depending on the aim of the model approach, it can be beneficial to combine several objective functions to find reliable posterior simulations (Houska et al., 2014). While this is not a surprising or new result, the advantage of SPOTPY is, that it facilitates an easy comparison of currently eleven objective functions in a pre- and post-processing mode.

Table I.2. Capabilities of the different algorithms implemented in SPOTPY. Checked fields indicate positive answers, fields with brackets are partly positive. [a] Only true during warm-up/burn-in. [b] Only true up to the number of used chains/complexes. They are separated on different CPU cores.

	MC	LHS	MLE	MCMC	SCE-UA	SA	DE-MC$_Z$	ROPE
Suited to investigate parameter uncertainty	✓	✓		✓			✓	✓
Allows considering multiple objective functions	✓	✓						
Possible to test prior parameter distributions	✓		(✓)[a]	(✓)[a]	(✓)[a]		(✓)[a]	(✓)[a]
Default algorithm-settings are all-round suited	✓	✓	✓	✓				✓
Suited for parallel computing	✓	✓			(✓)[b]		(✓)[b]	✓
Algorithm learns during sampling			✓	✓	✓	✓	✓	✓

In general, the findings reveal that not every algorithm is suited for every parameter search problem. Even more, every algorithm has its advantages and disadvantages. Therefore, the overview in Table I.2 showing the main capabilities of the algorithms might help the end-user to select a suited and efficient algorithm, without the need to understand and test every possible optimization technique. The approximate Bayesian compute techniques MC and LHS are very well suited to calibrate the model on multiple outputs with different objective functions. Nevertheless, they are very inefficient in high parameter space, like shown in the Ackley case study. Contrasting, the Metropolis MCMC method can be very efficient. However, it has the disadvantage that it is not possible to be

used in parallel computing systems. DE-MC$_Z$ is suited to be used in parallel, but gets inefficient when too many chains need to converge. ROPE is fully parallelizable but the generation of the parameter space after each subset needs a long computation time. All implemented non-Bayesian techniques (MLE, SCE-UA and SA) search only for one optimal parameter set, which makes them in general more efficient than the Bayesian approaches, but the outcome is very dependent on the used objective function and the parameter space, which is why they have to be chosen carefully. Furthermore, SCE-UA and SA need a pre-testing of the algorithm settings. They should not be used, without an adaption to a specific parameter search problems. MLE can be used straightforward, but the user has a higher risk to get stuck in a local optima. Unfortunately, there is no perfect algorithm and no perfect objective function. It depends. In this regard, SPOTPY was developed to help users to find their specific optimal solution.

Conclusion

As a final aspect, we want to check, if our five defined criteria are met by SPOTPY. We conclude that SPOTPY is a broad package, combining several optimization approaches. We hope that it is helpful to users, as no other parameter estimation package provides such a wide range of implemented techniques and is so easy to use. Optimization experts can still accessed and adopted the complexity of the algorithms. Modularity is given as the entire package is coded in Python. The independency of SPOTPY makes it applicable to every model; in contrast to other packages, e.g. the presented toolbox of the SWAT (Yang et al., 2008). The scalability claim of SPOTPY is valid. The straightforward MPI support results in a nearly linear time boost when analyzing time-consuming model runs and is as easy as tipping: *parallel='mpi'*. Finally, the open-source accessibility of SPOTPY makes it available for everyone to every field of science, where parameter optimization is useful. We will maintain the code at least for the next two years and expand the functionality systematically. For instance, the most recent version comes along with a sensitivity analysis algorithm (FAST) and more possibilities to structure the simulation data in the database. Finally yet importantly, we welcome new contributors to share their results or to provide new ideas for features.

Acknowledgements. The authors thank their colleagues for continuing support and discussion around the coffee breaks. We further thank Christoph Müller and Ludger Grünhage for providing the dataset of the FACE project.

II. Rejecting hydro-biogeochemical model structures by multi-criteria evaluation

This chapter is published in the journal "Environmental Modelling and Software" written by:

Houska, T.[1], Kraft, P.[1], Liebermann, R.[1], Klatt, S.[2], Kraus, D.[2], Haas, E.[2], Santabárbara, I.[2], Kiese, R.[2], Butterbach-Bahl, K.[2], Müller, C.[3,4] and Breuer, L.[1,5]: Rejecting hydro-biogeochemical model structures by multi-criteria evaluation, Environ. Model. Softw., 93, 1–12, doi:10.1016/j.envsoft.2017.03.005, 2017.

[1] Institute for Landscape Ecology and Resources Management (ILR), Research Centre for BioSystems, Land Use and Nutrition (IFZ), Justus Liebig University, Giessen, Germany
[2] Institute of Meteorology and Climate Research - Atmospheric Environmental Research (IMK-IFU), Garmisch-Partenkirchen, Germany
[3] Institute for Plant Ecology, Research Centre for BioSystems, Land Use and Nutrition (IFZ), Justus Liebig University, Giessen, Germany
[4] School of Biology and Environmental Science and Earth Institute, UC Dublin, Belfield, Ireland
[5] Centre for International Development and Environmental Research (ZEU), Justus Liebig University, Giessen, Germany

Abstract. This work presents a novel way for assessing and comparing different hydro-biogeochemical model structures and their performances. We used the LandscapeDNDC modelling framework to set up four models of different complexity, considering two soil-biogeochemical and two hydrological modules. The performance of each model combination was assessed using long-term (8 years) data and applying different thresholds, considering multiple criteria and objective functions. Our results show that each model combination had its strength for particular criteria. However, only 0.01% of all model runs passed the complete rejectionist framework. In contrast, our comparatively applied assessments of single thresholds, as frequently used in other studies, lead to a much higher acceptance rate of 40 to 70%. Therefore, our study indicates that models can be right for the wrong reasons, i.e., matching GHG emissions while at the same time failing to simulate other criteria such as soil moisture or plant biomass dynamics.

Introduction

The main anthropogenic source of N_2O is linked to emissions from agricultural soils and vast application of organic and synthetic nitrogen fertilizers (Reay et al., 2012). The underlying processes of soil carbon (C) and nitrogen (N) cycling and emission are affected by a multitude of non-linear factors, e.g. fertilization, tillage, climate, nutrient use efficiency as well as microbial metabolism (Stehfest and Bouwman, 2006). Consequently, greenhouse gas (GHG) emissions are highly variable in space and time. This variability across spatio-temporal scales cannot be addressed by field measurement as the spatial scale is too limited (Butterbach-Bahl et al., 2013). To overcome these limitations process based models, which summarize and translate our current understanding of processes underlying the biosphere-atmosphere GHG exchange into numerical equations, have been developed. These models allow upscaling in space and time domains and they can also be applied in the framework of scenario studies and used for decision support (Wang and Chen, 2012). Nevertheless, the algorithms used in such models are simplifications and still associated with

uncertainty since magnitude and parameterisation of many biogeochemical processes are uncertain, too (Butterbach-Bahl et al., 2013; Kraus et al., 2015).

Studies about biogeochemical model uncertainty analysis of GHG exchange processes and fluxes are still limited and differ with respect to implemented process descriptions, output targets and uncertainty sources. Lehuger et al. (2009) presented the first uncertainty analysis for a process-based biogeochemical model (CERES-EGC, a biogeochemical extension of the CERES crop model). The model output of N_2O fluxes were generated on 7 different sites with a Metropolis-Hasting algorithm, involving 15 model parameters. They found posterior model outputs with an uncertainty ranging from 13 up to 1422% for annual N_2O flux predictions. A review by Wang and Chen (2012) summarizes the few existing parametrization and uncertainty studies for soil biogeochemical models and recommend uncertainty analysis for multiple sites and the use of multiple criteria. They further suggest a development of a model library containing various model structures to facilitate comprehensive model comparison and uncertainty studies. Such a variable model structure approach was realized by Haas et al. (2013) with LandscapeDNDC (DeNitrification-DeComposition), a framework for simulation of water, C and N cycling and associate GHG emissions in terrestrial (forest, arable, grassland) ecosystems. LandscapeDNDC consists of interchangeable modules representing soil biogeochemistry e.g., scDNDC (Zhang et al., 2015) or the $MeTr^x$ module (Kraus et al., 2015), hydrology e.g., water cycle wcDNDC or CMF (Catchment Modelling Framework; Kraft et al., 2011), vegetation and microclimate processes. C and N turnover and related soil GHG emissions are, beside the main microbiological processes, depending on soil moisture conditions (Breuer et al., 2002; Butterbach-Bahl and Dannenmann, 2011). Consequently, to achieve reliable simulation of GHG emissions, an accurate representation of the soil moisture is a key requirement (Butterbach-Bahl et al., 2013; Frolking et al., 1998; Kröbel et al., 2010). Nevertheless, in biogeochemical models soil hydrological processes are often simulated based on simple bucket approaches, i.e. water moves vertically down a profile once a certain threshold has been reached, as e.g. in the LandscapeDNDC hydrological module wcDNDC. The nonlinear partial differential Richards' equation brings the advantage of a physical based approach. The equation describes vertical unsaturated flow, capillary rise and interaction with groundwater level. The implementation has been undertaken by Haas et al. (2013) and Wlotzka et al. (2014). They tested the coupled model system for C and N cycling on virtual hillslope studies including lateral nutrient transport. For sound validation of models, simulations must be tested with various observed data representing C, N and water cycling. However, most studies investigating biogeochemical processes and associated GHG emission simulated by the DNDC model family concentrate only on the validation of a subset of model results. Studies have been published with outputs such as N_2O emissions, yields or soil

temperature and moisture profiles (see literature survey of Giltrap et al. 2010) with the risk that simulated GHG emissions are right for the wrong reasons. To overcome this problem model testing should be done by taking as many different observations into account as possible. They further should be accompanied by an uncertainty analysis (Pappenberger and Beven, 2006). There is an intensive discussion about different sorts of model uncertainties and how to address them (Beven, 2015). One of the most widely used concepts for assessing model uncertainties is the Generalized Likelihood Uncertainty Estimation (GLUE) (Beven and Binley, 1992; Beven and Freer, 2001). GLUE has its origin in hydrological research but has been utilized in other scientific fields such as biogeochemistry or plant growth studies (Houska et al., 2014; Nylinder et al., 2011; Senapati et al., 2016; Wang and Chen, 2012).

In this study, we are interested in the benefits of a physically based process description over a conceptual approach in simulating soil water dynamics within a biogeochemical model. We follow the philosophy that complex models should be identifiable (low parameter uncertainty) and accurate (good agreement with observation data). Further, a model should be able to simulate various observation data concurrently and close to reality, especially when dealing with highly non-linear process interactions like in hydro-biogeochemical systems. To asses only such model runs, we perform a multi-criteria evaluation of different model structures and quantify their underlying uncertainties. This study combines the following points:

(1) We utilized a comprehensive, high quality, long-term data-set from a grassland study site in Linden, Germany (Jäger et al., 2003), which was established in 1998. Data of trace gas emission (N_2O, CO_2, cumulative CO_2 and N_2O), plant growth (biomass, cumulative biomass), and soil hydrology (soil moisture) was taken to evaluate the models.

(2) We established four model structures, by combining two varieties of the LandscapeDNDC biogeochemical modules with two soil moisture routines, resulting in the four model set-ups scDNDC/wcDNDC, scDNDC/CMF, MeTr^x/wcDNDC and MeTr^x/CMF.

(3) We reduced the parameter space of modules involved in GHG emission processes (e.g. decomposition, ammonification, nitrification and denitrification) through a stepwise sensitivity analysis.

(4) We run a multi-criteria GLUE for each model combination to find behavioural parameter sets and select appropriate model structures based on this assessment. Formally, we use a posteriori model rejection framework by selecting only those model structures that meet predefined objective functions (Vaché et al., 2004). The method is designed to detect and locate potential model and measurement errors. Our accepted model runs pinpoint such errors and help to analyse the data.

Methods

Model description

LandscapeDNDC is a simulation framework for terrestrial ecosystem models (Grote et al., 2009; Haas et al., 2013) with a modular structure allowing the easy and efficient combination and coupling of different modules describing different processes in ecosystem compartments, i.e. mathematical descriptions of microclimate, water cycle, plant physiology and soil biogeochemical processes. The modules are an abstract representation of the ecosystem. LandscapeDNDC defines six ecosystem compartment: canopy air chemistry, canopy and soil microclimate, vegetation physiology, vegetation structure (only for forest applications), water cycle and soil biogeochemistry. Every of this ecosystem compartments is represented by different modules, see Table II.1 and Figure II.1 for details. In this study, we test different combinations of two soil biogeochemistry modules (scDNDC and MeTr^x) and two water cycle modules (wcDNDC) and CMF in order to quantify model structure related uncertainty and to test validity of model structures. The different module combinations result in four model set-ups of the LandscapeDNDC framework, which are in the following referred to as scDNDC/wcDNDC, scDNDC/CMF, MeTr^x/wcDNDC and MeTr^x/CMF. For the plant physiology and microclimate, for all model set-ups we selected grasslandDNDC (Molina-Herrera et al., 2016) and canopyecm (Grote et al., 2009), respectively. All input settings for site and vegetation characteristics as well as climatic drivers are the same for any of the four tested models.

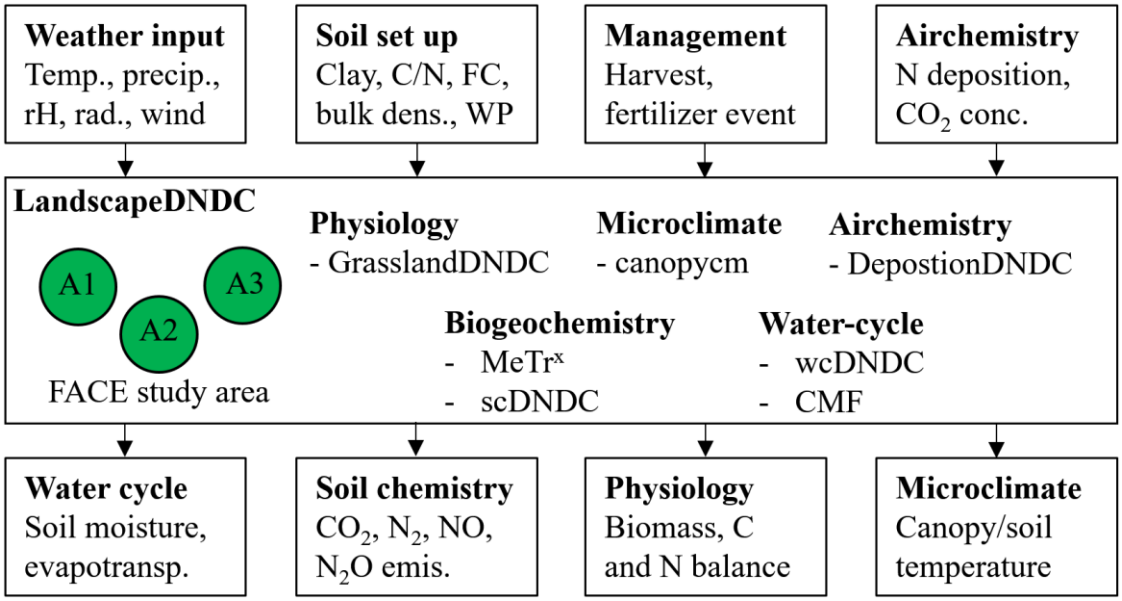

Figure II.1. Overview about the different inputs (upper boxes), modules (middle panel) and outputs (lower boxes) used in LandscapeDNDC for this study. For a complete input and output list of LandscapeDNDC see Haas et al. (2013). Temp. = temperature, precip. = precipitation, rH = relative humidity, rad. = solar radiation, FC = field capacity, bulk dens. = bulk density, WP = wilting point, evapotransp. = evapotranspiration, emis. = emission.

scDNDC

The scDNDC module originated from descriptions of the agricultural DNDC (Li et al., 1992) and ForestDNDC model (Li et al., 2000) and is suited for biogeochemical simulations of grassland, arable and forest soils (Haas et al., 2013). It addresses biogeochemical turnover of major soil C and N compounds. Turnover of soil organic matter and plant debris is largely controlled by C/N ratios and climatic factors, i.e. temperature and soil moisture. The focus of scDNDC is on the simulation of CO_2, N_2O and NO emissions (Butterbach-Bahl et al., 2009; Kim et al., 2015) as well as NO_3^-leaching (Dirnböck et al., 2016; Kiese et al., 2011). Production and consumption of C and N substrates are calculated based on microbial processes and metabolisms, e.g. mineralization, nitrification and denitrification, following the approaches of Blagodatsky and Richter (1998) and Leffelaar and Wessel (1988).

$MeTr^x$

$MeTr^x$ is a newly developed soil biogeochemical module for LandscapeDNDC (Kraus et al., 2015). In addition to C/N ratio and climatic factors, as in the case of scDNDC, decomposition of plant debris depends on its lignin and cellulose content following the concept of Bruijn and Butterbach-Bahl (2009) as well as the anaerobicity of the soil. Moreover, $MeTr^x$ uses a simplified formulation of denitrification with a reduced number of parameters (Bruijn et al., 2011). $MeTr^x$ has been successfully applied for the estimation of CO_2, CH_4 and N_2O emissions from tropical agricultural lowland and upland rice based cropping systems (Kraus et al., 2015, 2016).

wcDNDC

The water cycle module (wcDNDC) in LandscapeDNDC simulates evapotranspiration and soil water percolation depending on climatic input information, i.e., rainfall and temperature. Potential evapotranspiration is simulated using a modified Thornthwaite approach (Thornthwaite et al., 1957), while actual transpiration is calculated based on gross primary production and plant type specific water use efficiency. Soil water percolation is estimated based on a simple tipping bucket approach, mainly depending on physical soil characteristics, i.e. wilting point, field capacity and saturated hydraulic conductivity. A detailed description of the soil hydrology module is given by Kiese et al. (2011).

CMF

The Catchment Modelling Framework (CMF), developed by Kraft et al. (2011), is designed as a toolbox to build hydrological models (Houska et al., 2014). CMF was coupled with LandscapeDNDC (Haas et al., 2013) to simulate the water cycle in a more process-based manner. Potential evapotranspiration is calculated with the Penman-Monteith equation. Soil water percolation is

estimated with the Richards equation, matrix infiltration and the van Genuchten retention curve. Snow and ice accumulation as well as melting are simulated. Additionally solute transport of DOC and NO_3 are performed. This study uses CMF as a module within LandscapeDNDC. Settings of CMF can be adapted by setting the parameters for the water retention curve and hydraulic conductivity following the approaches van Genuchten (1980) and Mualem (1976), i.e. alpha, n and saturated hydraulic conductivity. Porosity is determined based on soil organic carbon and bulk density.

Model testing

In order to test the four different model set-ups and quantify model structure related uncertainty, we first applied a sensitivity analysis followed by a Generalized Likelihood Uncertainty Estimation (GLUE) method. For a meaningful evaluation of the performance by the different model set-ups, we use a field data set with comprehensive information on temporal dynamics of soil moisture and GHG exchange.

Field data and model set-up

For model testing, we used data from a well-studied extensively managed grassland site located at Linden, nearby Giessen, Germany. It is running since May 1998 consisting of three plots A1, A2, and A3, which were used in this modelling study. Mean annual precipitation and average temperature are 616 mm and 9.5°C, respectively. The vegetation and soil are characterized as *Arrhenatheretum elatioris* and fluvic gleysol (Kammann et al., 2008). GHG emissions of each plot are measured weekly with opaque static chambers (0.3 m height, 0.184 m^3 volume), sealed for 60-90 min to a soil collar and sampled in four 20-30 min time intervals (Kammann et al., 2008). Samples were analysed within 24 h for CO_2 and N_2O content with a gas chromatograph (HP6890) and GHG fluxes are determined, according to Kammann et al. (2008). Corresponding to the dark chamber measurement method, where a lightproof chamber is placed over the plants on the soil (see Kammann et al. (2008) for details), the measured soil CO_2 emissions were compared to the sum of simulated heterotrophic and autotrophic maintenance respiration of the plants. They reflect the respiration of the soil. Simulated autotrophic growth respiration was excluded assuming photosynthesis stops with chamber closure. This provides a way to model the measured plant-physiology darkness. Volumetric soil moisture of each plot is measured on working days with TDR sensors in 0-10 cm depth (Kammann et al., 2008). The vegetation is harvested in June and September each year 4 cm above the soil surface and fertilized in April with granular mineral calcium-ammonium nitrate (Kammann et al., 2008). We included all available site information (Table II.1) and calibrated only those model inputs where information was not available or very uncertain. Some soil properties change throughout the soil and measured values

increase (bulk density), decrease (clay content) or vary (pH) in depth. Soil organic carbon (SOC at topsoil) was implemented during model application by assuming an exponential decreasing to the lowest soil layers (in 0.55 m depth). All other soil properties (wilting point, field capacity, saturated hydraulic conductivity, van_Genuchten_alpha and van_Genuchtuen_n) were implemented by setting for the highest and lowest layer as parameters for GLUE and assuming a linear regression between the layers (Table A.1).

Table II.1. Input settings of the LandscapeDNDC model for the Linden grassland study site. In case spans are given, they reflect observed ranges for measurements FACE plots A1, A2 and A3, used throughout the set-up of the soil profile. Further detailed information on specific soil parameter values as used for the uncertainty assessment are given in Table A.1.

Input category	Input	Value	Unit	Source
Climate /Atmosphere	Daily temperature	9.5	avg. °C	Field observations
	Daily precipitation	616.2	mm a^{-1}	Field observations
	Annual N deposition	14.4	kg N ha^{-1}	Field observations
	Atmospheric background CO_2	402	ppm	Jäger et al. (2003)
Soil	Bulk density	1.01-1.52	g cm^{-3}	Jäger et al. (2003)
	pH	5.4-6.0	-	Jäger et al. (2003)
	C/N at surface	10	-	Kammann et al. (2008)
	Clay content	19-26	%	Field observations
Vegetation	Vegetation type	Perennial grass	-	Field observations
	C/N ratio above ground biomass	25.7	-	Field observations
Management	Fertilizer type	NH_4NO_3	-	Kammann et al, (2008)
	Annual fertilizer amount	40	kg N ha^{-1}	Kammann et al. (2008)

Model structure and parameter uncertainty

The model structure and parameter uncertainty of the four model set-ups was determined in two consecutive steps. In a first step, we identified the most sensitive model parameters affecting outputs of the biogeochemical, plant physiological and water cycle modules. These parameters were used in the following step for analysis of parametric uncertainty.

1) Sensitivity analysis of model outputs is widely used to determine the influence of a model parameter on a given output. We applied a variance based sensitivity approach, the Fourier amplitude sensitivity test (FAST) (Pianosi et al., 2016; Saltelli et al., 1999) to rank the more than 100 parameters of both soil biogeochemical modules, scDNDC and $MeTr^x$, according to their influence on the model outputs. Finally, we selected the 30 most sensitive parameters in order to come up with a reasonable number of parameters for the uncertainty analysis. For each biogeochemical module, the FAST algorithm repeated 100,000 iterations with different parameter combinations. The number of input parameters and their boundaries for the water cycle modules and physiology module were selected by expert knowledge. This resulted in 45 parameters for each model set-up (30 parameters of the soil biogeochemical module selected through the sensitivity analysis, plus four parameters of the water cycle module and eleven plant growth parameters of the physiology module).

2) Uncertainty analyses are designed to quantify the uncertainty of model predictions. The Generalized Likelihood Uncertainty Estimation method (GLUE, developed by Beven and Binley, 1992) is widely used to derive model uncertainty, by randomly sampling parameter combinations and accepting only ones fulfilling prior defined objective function criteria. Accepted model runs are defined as behavioural and similar good, also referred to being equifinal. We conducted a Latin hypercube sampling (LHS) with 100,000 repetitions for each of the four model set-ups, with each 45 parameters. See Table A.1 for the parameter names and their uniform priors. Detailed information about the meaning of these parameters can be found in the online documentation (LandscapeDNDC, 2015).

We used two objective functions to assess the model performances. For each model run, we quantified the Root Mean Squared Error (RMSE) and the mean bias (BIAS, also known as mBIAS). The RMSE is calculated according to Eq. (1):

$$F_1(\theta) = RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}, \quad (1)$$

where $F(\theta)$ is the objective function value for each model run given a parameter set θ, n is the total number of measurements, O_i is the observed value for the i-th measurement and P_i is the corresponding predicted value of the model. The RMSE can vary from 0 (perfect fit), towards large values (deep disagreement). This objective function has the benefit that it has the same dimension as the simulation values, which is helpful for quantification of the remaining uncertainties of the posterior simulations. The RMSE is further suitable for evaluating the modelled land term data agreement (Chang et al., 2013). See Table II.2 for the thresholds of the RMSE we set for behavioural model runs. These thresholds are oriented on the means of our observed data and other previous reported findings in grassland modelling studies (Chang et al., 2013; Lehuger et al., 2009). The threshold for soil moisture simulations is based on the mean standard deviation of the observed data. We calculated the BIAS as a second objective function. The BIAS is well suited to detect inter-annual differences and to assess whether structural changes of the model equations are necessary (Wallach, 2006). Yearly cumulative simulations were compared with yearly cumulative observations for every of the eight simulation year. The BIAS is calculated according to Eq. (2):

$$F_2(\theta) = BIAS = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i), \quad (2)$$

The BIAS can vary from 0 (unbiased estimator) to large positive or negative values (large deviation). We evaluated this objective function as a threshold for the cumulative GHG emissions as well as the cumulative biomass simulations for every year (Table II.2). Cumulative values of simulated GHG

emissions and biomass are commonly evaluated with overall monthly or yearly measured values (Kiese and Butterbach-Bahl, 2002; Zhang et al., 2015).

Using two different objective functions we assure on the one hand that the daily variation of GHG emissions and soil moisture (e.g. driven by management and/ or rainfall events) are reproducible by the model. The RMSE allows evaluating the ability of the model in reproducing daily variations; on the other hand, the BIAS permits to quantify systematic model error that can cumulate.

When we selected our thresholds, we tried to achieve a trade-off. Firstly, we oriented on previous reported findings, to achieve comparatively good results. Secondly, we found the mean of our observed data useful to account for site-specific characteristics. The finally selected thresholds are always in between these two boundaries:

N_2O emission: Our RMSE threshold of 0.0035 kg N-N_2O ha^{-1} d^{-1} is slightly lower than mean posterior RMSEs of seven study-sites in northern France (0.0055 kg N-N_2O ha^{-1} d^{-1}), derived with MCMC (Lehuger et al., 2009). Our threshold is slightly higher than mean measured fluxes with 0.0027 kg N-N_2O ha^{-1} d^{-1}, allowing model errors of >100%. We see this threshold as very permissive, due to still missing robust modelling concepts to simulate N_2O emission data.

CO_2 emission: Our RMSE threshold of 15 is slightly higher than mean RMSEs found in a modelling study for grasslands across Europe (8 kg C ha^{-1} day^{-1} published by Chang et al. (2013)). Our threshold is lower than the mean measured fluxes of 25 kg C ha^{-1} day^{-1}.

Biomass: Our RMSE threshold of 3,000 kg DM ha^{-1} is higher than other findings on an extensive grassland (1,800 kg DM ha^{-1}), derived with the plant growth model PROGRASS (Lehuger et al., 2009), but lower than our mean observed yields of 3,500 kg DM ha^{-1}.

Soil moisture: Reported calibrated RMSEs for WFPS can be found in a wide range, e.g. from 1.02% (Thorburn et al., 2010) to 24% (Ludwig et al., 2011), depending amongst other things on the variance of the measured soil moisture data. We decided to use the mean standard deviation (9.5%) across our plots A1-A3 as a comparatively demanding threshold.

The thresholds of BIAS are based on observed inter-annual variabilities. However, the selection of the thresholds remains subjective and other thresholds as well as other objective functions are thinkable, which would affect the results. We tried to reduce the methodological subjectivity by selecting strict thresholds and a high number of model repetitions, as recommended by Li et al. (2010). We have tested the acceptance criteria *a priori* to assure that they are achievable for the model. Overall, our approach results in 84 individual thresholds (4 whole-time outputs x 3 plots (A1, A2 and A3) with RMSE + 3 cumulative outputs x 8 years x 3 plots (A1, A2 and A3) with BIAS), see Table II.2. For the sensitivity analysis and the LHS based GLUE methodology we used the open source Statistical Parameter Optimization Tool in Python (SPOTPY) (Houska et al., 2015) on a High

Performance Computing system. We assessed all model structures matching a single criterion, but also looked for model structures that were able to meet objective function thresholds for all criteria (84-multi-threshold-filter). The percentage of remaining model runs is not necessarily better or worse. It does not give information about accuracy, only (indirectly) about parameter uncertainty. However, a model structure is likely to fail these constraints only if it is demonstrating a fundamentally different behaviour to reality, which is why a high number of remaining runs is good. This is obviously the opposite to a more common case where the thresholds are very demanding, such that many runs will remain only if there is an identifiability problem, which is bad. We are not aware of any study were this distinction has been noted before. To our knowledge, we are also the first conducting a multi-model, multi-criteria and multi-objective uncertainty assessment for a set of complex hydro-biochemical models.

Table II.2. Thresholds for a simulation with a distinct parameter set to be accepted as a behavioural model run.

Criteria	Objective function threshold			Unit	Evaluated for
Soil moisture	RMSE	<	9.5	vol.%	3 Plots
CO_2 emissions	RMSE	<	15	kg C ha^{-1} day^{-1}	3 Plots
N_2O emissions	RMSE	<	0.0035	kg N ha^{-1} day^{-1}	3 Plots
Biomass	RMSE	<	3,000	kg DM ha^{-1}	3 Plots
Cum. CO_2 emissions	BIAS	< ±	350	kg C ha^{-1} a^{-1}	8 Years, 3 Plots
Cum. N_2O emissions	BIAS	< ±	0.4	kg N ha^{-1} a^{-1}	8 Years, 3 Plots
Cum. Biomass	BIAS	< ±	3,000	kg DM ha^{-1} a^{-1}	8 Years, 3 Plots
					Total 84 thresholds

Results

Model efficiencies

After the sensitivity analysis, we run GLUE for the 4 model combination with 45 parameters, and selected the behavioural model runs fulfilling our objective function thresholds (Table II.2). Figure II.2 compares the remaining number of behavioural simulation runs for all four models. The percentage of accepted behavioural simulation runs indicates the effectiveness of each model configuration to find a parameter set that complies with the seven single criteria and the 84-multi-threshold filter.

Soil moisture: Only 4.4% of all 400,000 model runs were accepted, indicating a strong threshold for the soil moisture predictions. Model set-ups with $MeTr^x$ and CMF were more efficient compared to those with scDNDC and wcDNDC in simulating soil moisture. 60.0% out of the accepted model runs are from $MeTr^x$/CMF and 23.1% from scDNDC/CMF, indicating both a superior performance of CMF based model structures and possible correlation or insensitivity of the selected CMF parameters. The wcDNDC module contributes only a small percentage of the remaining posterior model runs in fitting the single threshold criteria (13.1% by$MeTr^x$/wcDNDC and 3.7% by scDNDC/wcDNDC).

Soil CO_2 emissions: 36.8% of all model runs were accepted to reproduce the soil respiration data, with a share of 85.0% of these simulations having MeTr^x as a soil biogeochemical module and 41.8% with CMF as the water cycle module. For the cumulative soil respiration, only 7.4% of all model runs were accepted. Here, MeTr^x model set-ups contribute to even 95.6% of model runs and CMF-based models to 45.3%. While we found a clear preference for the MeTr^x soil chemistry module, wcDNDC was only slightly more represented than CMF.

Soil N_2O emissions: 68.6% of all model runs were accepted, indicating both a relatively week threshold and the suitability of DNDC to simulate N_2O fluxes, something the model was initially developed for. Out of these, the scDNDC module is most efficient in simulating N_2O emissions and contributes 68.6% of all accepted simulations. There was no clear preference on the soil moisture routine, for which CMF contributed 52.0% and wcDNDC 48.0%, respectively. The pattern of contributing modules for cumulative N_2O emissions was similar, though less model runs were accepted (59.4%) and scDNDC/wcDNDC outperformed the other module combinations with almost contributing 39.7% of all runs.

Biomass: More than half of the model runs were accepted, with equal contributions for both wcDNDC set-ups and the combination of MeTr^x/CMF. Similar to CO_2 emission, scDNDC/CMF contributes the lowest ratio of 0.6%. The threshold for cumulative biomass harvest was met by 7.8% of all behavioural model runs, from which 94.8% where MeTr^x based models, with MeTr^x/wcDNDC slightly outperforming MeTr^x/CMF. We found no acceptable model set-up for the scDNDC/CMF model.

Multi-threshold criterion: Fulfilling all 84 thresholds was only achieved by 0.01% of the model runs; 12 MeTr^x/wcDNDC, 11 MeTr^x/CMF and 8 scDNDC/wcDNDC model configurations. Only these model set-ups and parameterizations are well identified. They performed equally well and are further defined as behavioural. The scDNDC/CMF model configuration was not able to fulfil the multi threshold criterions. The accepted model set-ups were defined as the posterior model runs, and were investigated in more detail in the following sections.

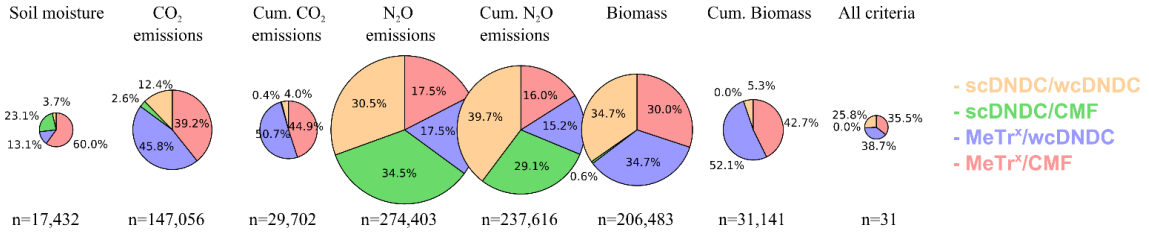

Figure II.2. Ratio of different model set-ups that fulfil single criterion threshold. The number of accepted model runs is given by *n* and is illustrated by the size of the pie chart (number of total model runs: 400,000). Only those model runs are depicted that fulfil the objective function thresholds given in Table II.2.

Model performance – analysis across plots

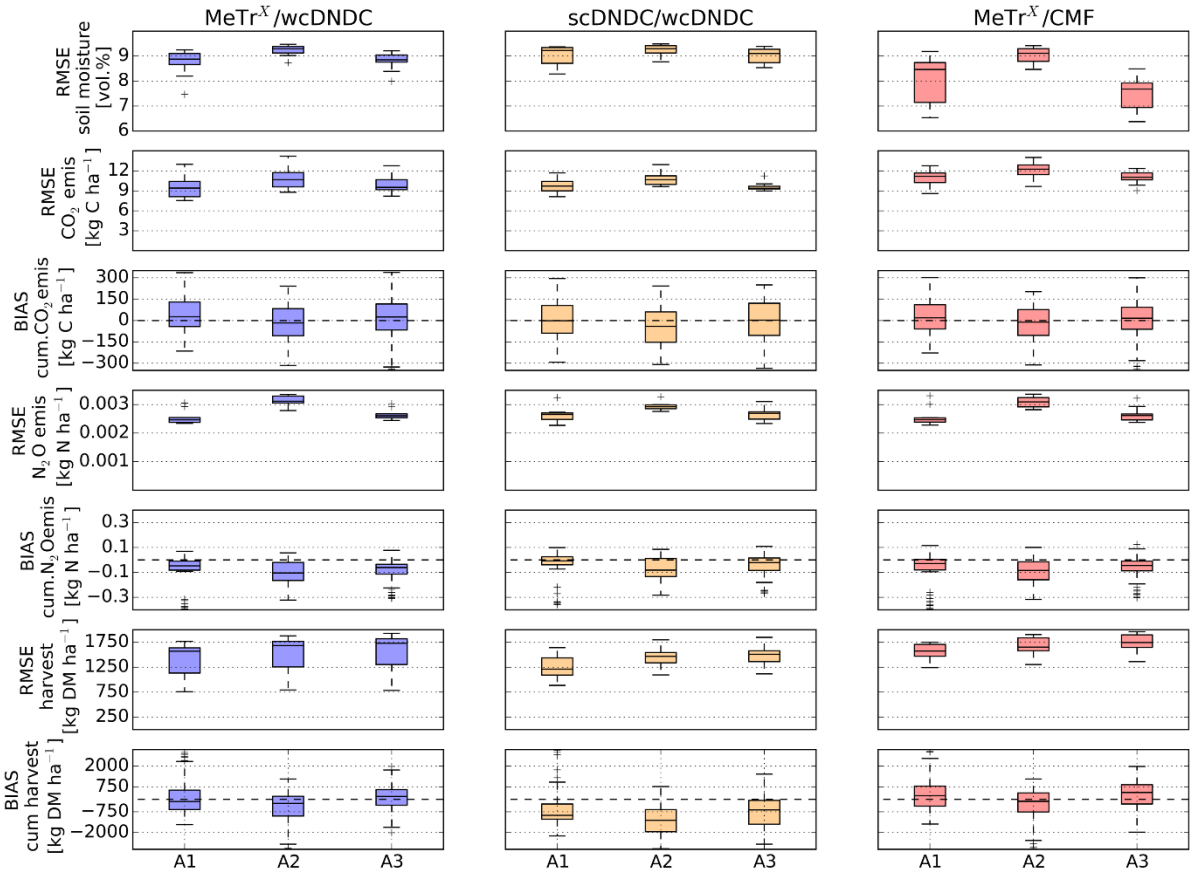

Figure II.3. Distribution of the objective functions of the posterior model runs selected by the 84-multi-threshold-filter for each of the three retaining model structures. Black dotted line is the optimal value for the BIAS. Boxes indicate the 25 and 75 percentile, whiskers the 5 and 95 percentiles, the median is the bold line in the boxes and outliers are shown as pluses. A1, A2 and A3 are the three different plots of the investigated grassland site. emis. = emission, cum. = cumulated.

For the analysis of the model performance, we used the objective functions of the posterior model runs as indicators. Figure II.3 shows the distribution of the posterior objective functions for the 31 accepted simulations for each of the three plots.

The observed soil moisture data can be simulated by the models with a rather high RMSE of 6 to 9.5 vol.%. Compared with the other model structures, the MeTr^x/CMF shows the best performance in terms of soil moisture, especially for plots A1 and A3, where some of the RMSEs are about 1 to 2% lower than for the other model set-ups. The RMSEs of the soil respiration vary between 8 to 13 kg C ha^{-1} day^{-1}, with MeTr^x/wcDNDC performing slightly better compared to the other model combinations. However, the spread of simulations for this model structure shows a larger output uncertainty reflected by the wider box plots. The posterior objective functions of the cumulative CO_2 emissions are normally distributed and all models performed similar on all plots. The N_2O emissions

of plots A1 and A3 can be simulated with an RMSE of around 0.0025 kg N ha^{-1} day^{-1}, which represents 92% of the overall mean N_2O emissions. Simulations for A2 are slightly worse with medians of 0.0030 kg N ha^{-1} day^{-1} (111% deviation). scDNDC/wcDNDC reproduced the N_2O emissions on A2 slightly better than the configurations using the $MeTr^x$ biogeochemistry. The cumulative N_2O emissions were in tendency underestimated by all models by 0 - 700%, indicated by a negative BIAS. The model scDNDC/wcDNDC was again slightly better in simulating A2 and A1 than the other models. The negative bias, especially on site A2 indicates a substantial difference in the measured data between the plots A1-3. Figure A1 indicates that single high measurements for N_2O can already have a large influence on the annual emissions. Measurements errors can easily effect the model results at this point. The RMSE for biomass harvest simulations vary from 700 to 2,000 kg DM ha^{-1} (average observed yield 3,493 kg DM ha^{-1}). We found the highest variation for $MeTr^x$/wcDNDC model, while the scDNDC/wcDNDC model resulted in the lowest RMSEs. Nevertheless, the scDNDC/wcDNDC model underestimated the cumulative biomass harvest depicting relative larger underestimations (median of -500 to -1,000 kg DM ha^{-1}). Here the $MeTr^x$ models performed better, especially on plot A2.

Model performance – analysis in time

In order to investigate the performances of the model structures in time, we compared simulations for the three remaining model structures with the observations of the three plots (A1-3). Presented results (n=31) are derived with behavioural parameter sets, which passed the 84-multi-threshold filter. Figure II.4 illustrates the simulated daily patterns of soil moisture, trace gas emissions, and biomass production over the simulation period of eight years. Corresponding patterns of annual cumulative trace gas emissions and biomass production are given in Fig. A1.

Soil moisture: The $MeTr^x$/CMF model reproduced the daily measured data best, particularly during dry conditions (soil moisture < 30 vol.%). This is most obvious during dry spells for example in the summer of 2000 and 2001. Overall, the Richard's equation based model $MeTr^x$/CMF shows stronger variability in the soil moisture than the tipping bucket approach of wcDNDC. This uncertainty bounds of the model $MeTr^x$/CMF overlap on 57.3% of the simulation days with the measurement uncertainty, while for the wcDNDC this is only the case for 46% of the simulation days. The standard deviation of the measurements was generally higher (9.5%) than for the simulations (ranging from 3.2% for scDNDC/wcDNDC to 5.1% for $MeTr^x$/CMF).

Soil CO_2 emissions: The seasonal pattern of daily CO_2 emissions are predicted by all models with an average span of the lower and upper simulated bounds of 15 kg C ha^{-1} day^{-1}. The bounds in early winter (November until January) have a smaller span of 10 kg C ha^{-1} day^{-1} and is highest (up to

20 kg C ha^{-1} day^{-1}) in May to September Overall, the variation resulting from the simulations is slightly higher than the variation of the emission measurements at the plots A1-A3. There was no model that sufficiently reproduced peak emissions, e.g., after the fertilization event on 18th April 2000 or in summer 2002. We believe that we face measurement errors in this case and the process model will never be able to reproduce them. The approach helps to pinpoint such possible measurement errors and helps to analyse the data. The better performance of the $MeTr^x$/CMF model for soil moisture simulations, particularly during dry periods, did not result in improved predictions of soil respiration compared with the other two models. Average yearly measured soil respiration was around 7.8 t C ha^{-1} a^{-1} (note that Fig. A.1 is showing only cumulative values for measured days, which results in lower values).

Soil N_2O emissions: The daily N_2O emission dynamic was reproduced with a mean uncertainty band of 0.003 kg N ha^{-1} day^{-1} by the three model structures, differing from close to zero in winter months up to 0.086 kg N ha^{-1} day^{-1} on single events. scDNDC/wcDNDC reproduced the high emissions in the first simulation year of 1999 better than the other models, but generally overestimated the emissions after 2002. This period was much better matched by $MeTr^x$/wcDNDC and $MeTr^x$/CMF. During low soil moisture conditions, e.g. summer 2004, the $MeTr^x$/CMF simulated the N_2O emissions better than the two other models. Obvious for all models is the substantial underestimation of pulse emissions particularly at day 13th November 2000 when all models failed to predict an increase of N_2O emissions. Average yearly measured N_2O emissions were around 0.9 kg N ha^{-1} day^{-1}.

Biomass: The seasonal biomass production was simulated with an uncertainty of ± 2,000 kg DM ha^{-1} (57% of the mean observation data) for the different model structures. The observations showed a higher yield at the first cutting in May than at the second cutting in September. The models failed to follow this temporal pattern and showed in most of the simulations a lower yield in May than in September. In addition, the uncertainty of simulations at the second cutting event were two- to threefold higher than at the first cutting event. The resulting uncertainties (RMSE) between single harvest events varied between highest RMSE of 7,000 kg DM ha^{-1} for $MeTr^x$/wcDNDC and 1,000 kg DM ha^{-1} for $MeTr^x$/CMF. However, simulation deviations in yearly means were lower than 3,000 kg DM ha^{-1}. The simulated yearly cumulative biomass production decreased for all models over the 8 years simulation period. While in the first two years for the first cutting event uncertainty bounds of simulations are in the same range than observation uncertainties, there is a clear tendency to under predictions in the following years. However, summing up the harvested biomass of both cuts, the $MeTr^x$ based models seem to perform better than the scDNDC based model.

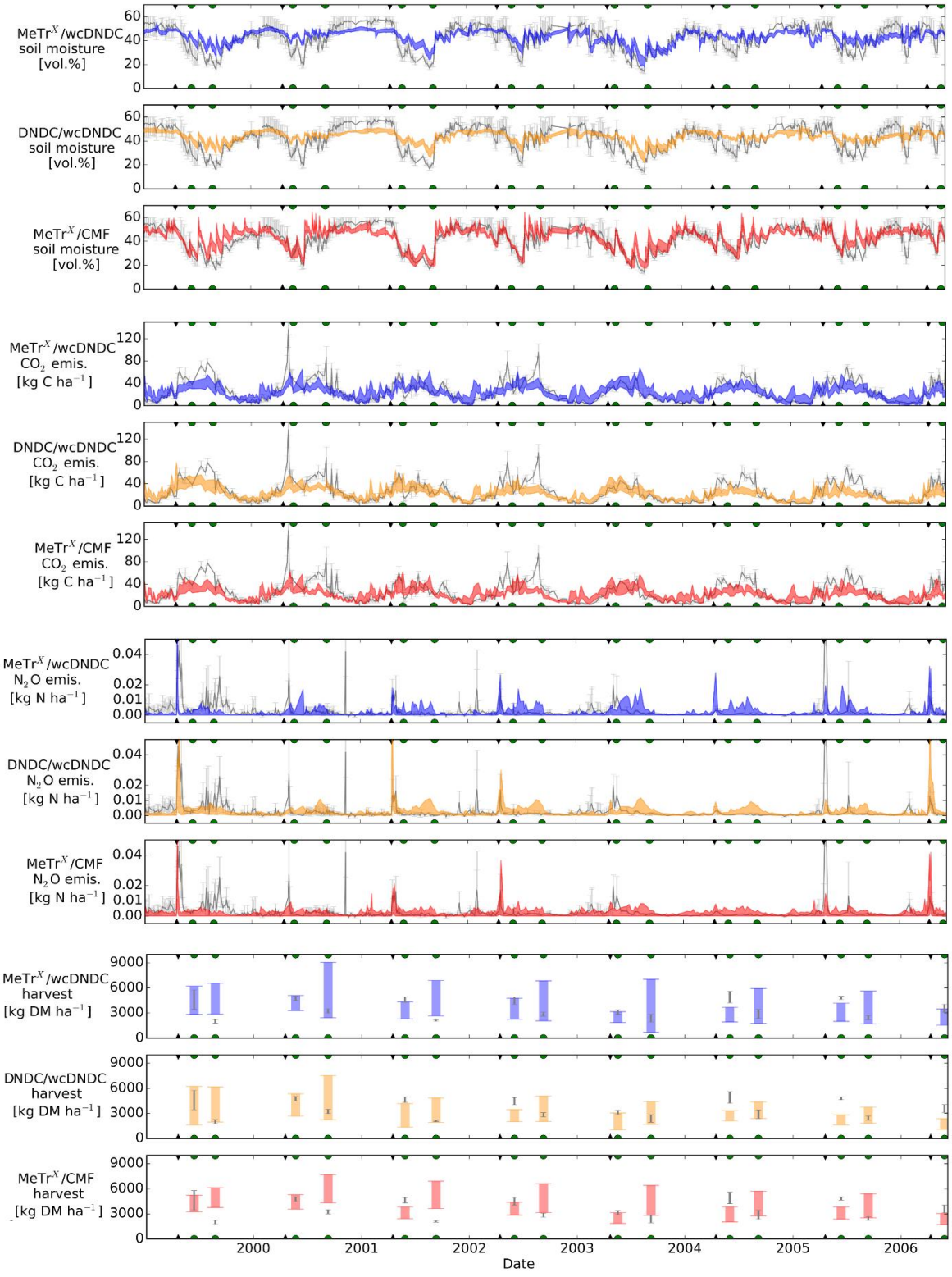

Figure II.4. Upper and lower bounds (coloured bands) of the posterior distribution model runs selected by the 84-multi-threshold-filter for each of the three retaining model structures. Grey lines of daily soil moisture and trace gas emissions as well of bars for biomass indicate mean observations ± standard deviation calculated from the single plots A1-3. Black triangles indicate fertilizer events, green semi-circles hay cutting.

Discussion

According to our selected single objective function thresholds, all four different model structures resulted in a sufficient number of behavioural model runs, some of them even exceeding 200,000 out of 400,000 randomly sampled parameter sets. In contrast to the single criterion, only 0.01% of all model setups passed the 84-multi-threshold filter. This strong reduction of the acceptance rate indicates that most of our model structures and parameterization that were selected based on the single criterion evaluation might be right for the wrong reason. Unfortunately, no matter how good models are initialized and run with measured input information (e.g. soil texture, field management or meteorological information), parameter and model structural uncertainties are likely to misdirect non-calibrated model results.

Here we provide evidence that complex, process-based hydro-biogeochemical models need to be thoroughly tested and checked against multiple-criteria using appropriate objective functions. For example, soil moisture is an important driver for many microbial C- and N-turnover processes and associated GHG emissions (Butterbach-Bahl and Dannenmann, 2011). Studies where daily observed and simulated soil moisture data are compared remain scarce and if presented, results are often not very satisfactory. A soil moisture simulation study with the tipping bucket approach of DNDC resulted in RMSE not better than 34.6% WFPS on a cropland in northern China (Cui et al., 2014) and another study using DNDC resulted in an RMSE of over 15% WFPS in the upper 10 cm of an soil under arable land use (Gao et al., 2014). Other studies like Lehuger et al. (2009) used a Markov Chain Monte Carlo algorithm to calibrate the CERES-EGC model on WFPS and N_2O across different sites. They found much lower RMSEs varying between 3 to 6% WFPS, by using a physical sub-module based on the generalized Darcy's law. In our study, the CMF representation of soil water transport by the Richards' equation was superior compared to the simple tipping bucket approach, however still far from a perfect prediction. It is well known that effective saturation calculated by Van Genuchten (1980) is subject to high uncertainty (estimation of parameters alpha and n by effective saturation data is difficult). However, any water transport model, including wcDNDC, needs to be characterized by some parameters describing water retention, e.g. soil moisture at wilting point and field capacity. In such simplified tipping bucket models, the parameter uncertainty is augmented with a higher structural uncertainty of the model. However, compared to other estimated parameters, optima of alpha, n and saturated conductivity were well identified (Figure A.10).

Likely reasons are in field heterogeneity of soil parameters that are used to either initialize or parameterize the DNDC modules or pedotransfer functions of CMF. Particularly pedotransfer functions can be extremely sensitive to input parameter variability (Bormann et al., 2007). Further,

soil moisture was taken as one mean value (out of 4 measurement points) within each plot, contributing to validation data uncertainty that we did not account for in this study. McMillan et al. (2012) showed this impact for a hydrological model, where they assessed the uncertainty of the stage-discharge relation on the error of the derived and simulated discharge. Simply comparing the number of soil moisture observations that are covered by the upper and lower bounds simulations gives a glimpse on this aspect: while the behavioural model runs of MeTrx/CMF captured 29.4% of the mean soil moisture, this value increased to 57.3% when the span of soil moisture observations was taken into account.

Emissions of N_2O are frequently simulated with biogeochemical models such as DAYCENT, ORCHIDEE, ECOSYS and are one of the primary targets of DNDC model applications (Gilhespy et al., 2014). Bouwman et al. (2010) state that many studies do not report day-to-day validation of N_2O emissions because of the poor performance of most models or the lack of available data. The deviations found in our study are in the range of best agreements between simulated and measured values reported from other studies in the literature. For DAYCENT applied on cropland RMSEs of 0.0024 kg N ha^{-1} day^{-1} (Rafique et al., 2013) and 0.108 kg N ha^{-1} day^{-1} (Necpálová et al., 2015) were reported for N_2O emissions. The uncertainty of our yearly simulated N_2O emissions is ±60% of the yearly measured emissions, which is within the measured uncertainty of ±124%. Lehuger et al. (2009) reported uncertainties of different sites with values ranging from -4 to 8% and uncertainties up to -68% to 154% derived with CERES-EGC. The reason for this high variation of the model performance in simulating N_2O emission are multiple sources, often explained by the "hole in the pipe" concept (Firestone and Davidson, 1989). N_2O emissions result from nitrification at medium WFPS in and denitrification at high WFPS. Although, models can account different N_2O production pathways, the validation is not possible due to still missing measurements of N_2O source partitioning. Consequently, here, we concentrate only on emissions during times of the year, which differ in their potential for N_2O emissions. High N_2O emissions are typically found after fertilization events (Liu and Greaver, 2009). The simulation of such events results in high remaining uncertainty, as we have shown for the scDNDC based models. Another hot moment of N_2O emissions are freeze-thaw events, which can significantly contribute to the yearly N_2O budget (Holst et al., 2008; Wolf et al., 2010), which is also true during strong winters (e.g. 2000/ 2001 and 2001/ 2002 Figure II.3) at our Linden grassland sites (Müller et al., 2002). Since impacts of freeze-thaw events on nutrient availability and soil microbes (de Bruijn et al., 2009) are currently not implemented in the applied biogeochemical models, elevated N_2O emissions during such events cannot be reproduced by any simulation. Consequently, if such events are important for the annual N_2O budget the model performance decreases.

A study with the model ORCHIDEE reports RMSEs of 5 to 20 kg C ha^{-1} day^{-1} for soil respiration on different grassland sites (Chang et al., 2013), which is a wider range than we found for our grassland sites. As a site aspect, we could show that the Richards equation based model did not only improved the soil moisture simulation, it reduced also the soil respiration uncertainty, during very wet and dry conditions. This effect can be explained by a strong positive correlation of soil moisture and CO_2 emissions from soils (Gong et al., 2014). Especially during dry conditions in the early summer months with lower above ground biomass, yield the $MeTr^x$/CMF model more accurate predictions of soil moisture also better predicts the CO_2 emissions, than the wcDNDC based model. Besides the soil moisture, CO_2 fluxes are depending on soil temperature, and re-wetting event. Especially the latter one is not yet accurate implemented in the model and might explain simulation days differing up to 100% from the measurements. Despite model failure one has to keep in mind that measurement of soil respiration are challenging and the fluxes, from in our case only three replicates can be associated with uncertainty.

We could fit the cumulative biomass over the year, but had difficulties in representing the seasonal trend of the grassland biomass with a BIAS of up to 85% difference. Other studies reported RMSEs of 1,300 to 1,800 kg DM ha^{-1} for grass mixture simulated with PROGRASS (Lazzarotto et al., 2009), depicting an overall better model performance than in our study. However, this model was particularly set-up for physiology based grassland biomass simulations, while the plant growth submodule used in our study follows more empirical based descriptions. In contrast to soil moisture content and GHG emission measurements, which are characterised by comparatively high uncertainties, biomass measurements are straightforward and robust. Thus, more effort should be put in the refinement and further development of the grassland plant growth module e.g. by splitting into plan functional types of grass, herbs and forbs instead of a homogeneous stand mixture (green leave approach) that is currently simulated. Moreover, nitrogen could be a limiting nutrient. Upwelling water or capillary rising water could deliver NO_3^- from lateral moving N containing groundwater into the grassland rooting zone and thereby promoting plant growth. This process, which can be of importance at the simulated site, is currently not implemented into the models. First field observations support this potentially important N source, with groundwater concentrations found in the range of 3 to 5 mg N l^{-1}. We summarized the detected model errors in Table II.3.

Table II.3. Summary table of remaining model weaknesses identified in this study.

Module	Detected module error	Proposed improvement
General	Missing freeze-thaw cycle	Include process e.g. based on Wolf et al. (2010).
	Failed to reproduce increased soil respiration after rewetting	Check process description in biogeochemical modules.
	Failed to reproduce hot moments	Include nitrate as input from groundwater.
	Failed to reproduce inter-annual biomass dynamic	Differentiate between types of grass, herbs and forbs instead of a homogeneous stand mixture.
MeTr^x	Parameter poorly identified	Reduce number of parameters and parameter range.
scDNDC	Fails with better soil moisture representation.	Check for structural errors and the coupling with CMF.
CMF	High remaining parameter uncertainty through Van-Genuchten retention curve	Test other retention curves and set stricter objective function thresholds.
wcDNDC	Failed to capture soil moisture dynamic	Set stricter objective function thresholds.

The strategy to address multiple observations is complex and presented herein simplified. One could claim that a low number of remaining model runs would be beneficial; indicating a better support through well-defined parameters, while a high number would indicate correlated parameters. However, our results indicate, that efficient model structures have also the best performing objective functions (compare Figure II.2 and II.3). The results could be further refined e.g. by performing a ranking of the behavioural model runs (Beven, 2006), by using more sophisticated optimization methods like DREAM (Vrugt et al., 2003), or by applying more theoretically based methods like the Iterative Closed Question Modelling (ICQM), introduced by Guillaume et al. (2015). Finding the underlying source of model malfunctioning is difficult, especially when using non-linear models and errors that might interact non-linear (Beven, 2007). However, we would like to raise awareness that there exists a surprisingly small intersection of equally well model performance on different model outputs, which leaves room for further analysis and model improvement.

Despite the questionable efficiency of the model structures to represent all observed data sets at the same time, the observation data itself are highly variable. This data has been measured on three plots with the same land use and soil type within a distance of less than 100 m on even topography. The daily soil moisture (37.2±9.0, 46.2±11.6, 40.1±11.2 vol.%) and weekly soil respiration (25.0±75.4, 26.1±78.3, 23.8±67.8 kg C ha^{-1} day^{-1}) data show significant differences between the plots. Possible influence of groundwater on the plots A1-3, which we cannot quantify yet, might explain part of the differences. Nevertheless, we face several question about data uncertainty and data quality. Are the available data sets sufficient for model evaluation? How do we deal with data uncertainty in larger scales? How can we improve the measurements? Are weekly trace gas measurements sufficient to capture all relevant processes? Future modelling studies can help to address these questions, by having a closer look on the process validation instead of just fitting models on observed data.

Conclusions

We conclude that process based models can only be as accurate as the current understanding of the system allows. While there are several ways to test models and their uncertainty, we showed one straightforward way to deal with some sources of uncertainties by using a GLUE based rejectionist framework. The replacement of the conceptual tipping bucket with the nonlinear Richards equation enables to reproduce temporal dynamics of soil moisture better. Furthermore, this approach reduces the simulation uncertainty of soil respiration. This point is linked to a better representation of reality, at the cost of a higher computational demand. The comprehensive data set of the grassland study site in Linden gave us the opportunity to validate different model configurations of LandscapeDNDC more intrinsically. While the representation of the N_2O emissions is in range with other studies, we saw the model struggling in reproducing the temporal pattern of the biomass data. We could show that multiple objective functions constrain the number of behavioural model runs much more than single objective functions can do. This point results in a dramatic drop of efficiency from acceptance rates up to 70% down to 0.01%. Our results stress that model outputs are not reliable, if they are not been tested against observation data. We need comprehensive field data for future hydro-biogeochemical modelling studies and a better understanding of the uncertainty in measurements. Upscaling simulations while ignoring these diverse uncertainties can easily misdirect our conclusions. We therefore believe that the communication of the modelling and measurement uncertainty should be part of good scientific practice in modelling studies (Pappenberger and Beven, 2006) as is also requested by the IPCC for the UNFCCC GHG reporting.

Acknowledgements. The authors thank their colleagues for continuing support and ongoing discussions. Part of this work was funded by the LOEWE excellence cluster FACE2FACE of the Hessen State Ministry of Higher Education, Research and the Arts. We acknowledge the financial support provided by the Deutsche Forschungsgemeinschaft (DFG) for Tobias Houska (BR2238/13-1) as well as Edwin Haas and Ignacio Santabarbara (BU1173/12-1). Edwin Haas received additional support from the FACCE-JPI MACSUR II.

III. Constraining of biogeochemical models with multi-site N_2O and CO_2 emission simulations by model-data fusion

This chapter under review for the journal "Biogeosciences" written by:

Houska, T.[1], Kraus, D.[2], Kiese, R.[2] and Breuer, L.[1,3]: Constraining a complex biogeochemical model for multi-site greenhouse gas emission simulations by model-data fusion, Biogeosciences Discuss, 2017, 1–28, doi:10.5194/bg-2017-96, 2017.

[1] Institute for Landscape Ecology and Resources Management (ILR), Research Centre for BioSystems, Land Use and Nutrition (IFZ), Justus Liebig University, Giessen, Germany
[2] Institute of Meteorology and Climate Research - Atmospheric Environmental Research (IMK-IFU), Garmisch-Partenkirchen, Germany
[3] Centre for International Development and Environmental Research (ZEU), Justus Liebig University, Giessen, Germany

Abstract. This study presents the results of a combined measurement and modelling strategy to analyse N_2O and CO_2 emissions from adjacent arable land, forest and grassland sites in Germany. The measured emissions reveal seasonal patterns and management effects, including fertilizer application, tillage, harvest and grazing. The measured annual N_2O fluxes are 4.5, 0.4 and 0.1 kg N ha^{-1} a^{-1}, and the CO_2 fluxes are 20.0, 12.2 and 3.0 t C ha^{-1} a^{-1} for the arable land, grassland and forest sites, respectively. An innovative model-data fusion concept based on a multi-criteria evaluation (soil moisture at different depths, yield, CO_2 and N_2O emissions) is used to rigorously test the LandscapeDNDC biogeochemical model. The model is run in a Latin Hypercube based uncertainty analysis framework to constrain model parameter uncertainty and derive behavioural model runs. The results indicate that the model is generally capable of predicting trace gas emissions, as evaluated with RMSE as the objective function. The model shows a reasonable performance in simulating the ecosystem C and N balances. The model-data fusion concept helps to detect remaining model errors, such as missing (e.g., freeze-thaw cycling) or incomplete model processes (e.g., respiration rates after harvest). This concept further elucidates the identification of missing model input sources (e.g., the uptake of N through shallow groundwater on grassland during the vegetation period) and uncertainty in the measured validation data (e.g., forest N_2O emissions in winter months). Guidance is provided to improve the model structure and field measurements to further advance landscape-scale model predictions.

Introduction

Carbon dioxide (CO_2) and nitrous oxide (N_2O) are two prominent greenhouse gases (GHG) contributing to global warming, the latter having a global warming potential (GWP) 300 times higher than that of CO_2 considering a 100-year time horizon (Myhre et al., 2013). Terrestrial ecosystems play an important role in the global atmospheric budgets of both GHGs (Cole et al., 1997). The global CO_2 emissions from soils are five times higher than anthropogenic (mainly fossil fuel) CO_2 emissions

(Raich and Schlesinger, 1992; updated with recent fossil fuel data by Boden, et al., 2010), while agricultural land use released over 60% of the global anthropogenic N_2O emissions in 2005 (IPCC, 2007). In addition to the radiative forcing of both GHGs, N_2O is currently the main driver of stratospheric ozone depletion (Ravishankara et al., 2009), causing increased ultraviolet radiation, which could result in skin cancer and other health problems (Graedel and Crutzen, 1989). While CO_2 is exchanged with the soil (heterotrophic respiration) and vegetation (photosynthesis and autotrophic respiration), N_2O fluxes refer mainly to the nitrification and denitrification processes occurring only in the soil (Butterbach-Bahl et al., 2013).

Emissions of both GHGs are highly variable in space and time and depend on a multitude of different interacting environmental factors, e.g., land use/management, nitrogen/carbon inputs, meteorological conditions and physical and chemical soil properties (Davidson, 1992; Smith et al., 2003). They are largely regulated by plant physiological (Rochette et al., 1999) and microbial processes (Burton et al., 2008). Field measurements of GHG emissions and environmental drivers have paved the way for a basic understanding of observed emissions patterns. Nevertheless, the large number and complexity of the processes involved in the production and consumption of CO_2 and N_2O are still challenges in the reliable quantification of related GHG emissions (Butterbach-Bahl et al., 2013). Various biogeochemical models have been developed in recent years. These models are used for temporal as well as spatial up-scaling of GHG emissions, hypothesis testing of our understanding of processes, and, for scenario analyses and the evaluation of efficient mitigation options (Kim et al., 2015; Molina-Herrera et al., 2016). These include, e.g., BASFOR (Oijen et al., 2005), CERES-EGC (Gabrielle et al., 2006), COUP (Jansson, 2012), DAYCENT (Parton et al., 1998) and DNDC and its descendant LandscapeDNDC (Haas et al., 2013). However, models are still simplifications of the real world and are prone to multiple sources of uncertainty, i.e., defective model structure and/or parameterization and the current model state (Vrugt, 2016). During model application, poor-quality model forcing data results in further uncertainties about the predicted model outcome (Kavetski et al., 2006). However, there is still no method available to properly address these sources of uncertainty at the same time (Vrugt, 2016). One promising way to reduce the magnitude of uncertainties in model output is to use model-data fusion techniques, i.e., matching model prediction with multiple observations by varying model parameters or states using statistical uncertainty estimation (Keenan et al., 2011). There are several statistical uncertainty estimation methods available, e.g., formal Bayesian approaches such as DREAM (Vrugt, 2016) and informal Bayesian approaches such as GLUE (Beven and Binley, 1992). However, these approaches are mostly used to fit models to single types of observations (Giltrap et al., 2010). Innovative multiple observation data evaluations with model-data fusion are becoming common in ecosystem carbon modelling (Wang et al., 2009) and are more and more important in the

nitrogen modelling community (Wang and Chen, 2012). The knowledge gained can and should be used to guide further model improvements (Vrugt, 2016).

This work focuses on establishing model-data fusion in the biogeochemical community – i.e., showing the capability of this technique to improve process understanding through the application of process-based models. We present weekly measurements of CO_2 and N_2O emissions from a developed landscape with different land uses, i.e., arable land, grassland and forest ecosystems, covering a two-year period of observations. In addition to field measurements, we set up the biogeochemical LandscapeDNDC model for each of the three land uses. During model-data fusion with GLUE, we rigorously accept only model runs that return concurrent, acceptable outputs for N_2O, CO_2, and soil moisture at different depths and yields. Posterior model runs are not only evaluated as to whether they fulfil appropriate objective functions but also regarding realistic simulations of GHG emissions for separate seasons, annual sums as well as before and after land management. The model is finally used to estimate the magnitude and uncertainty of C and N fluxes, such as N_2 emissions or autotrophic and heterotrophic CO_2 emissions, which are not yet experimentally quantifiable *in situ*. The remaining model and data errors are traced back to their potential sources to improve ongoing measurements and future model applications.

Material and methods

Study area

The study area is located in the catchment of a low mountainous creek (Vollnkirchener Bach) in the municipality of Hüttenberg, Hesse, Germany (50°29′56″ N, 8°33′2″ E). One kilometre north of the village of Vollnkirchen, next to the creek, we established eight transects (oriented mostly vertically to slope) along a valley cross-section covering different types of land uses (Figure III.1) for GHG emission measurements. See Table III.1 for detailed information on soils characteristics. Three transects (A1-A3) are located on arable land to the west of the creek and were cultivated with the same field management and crop rotations (Table III.1). Three transects are located in a light beech (*Fagus sylvatica*) forest (W1-W3) with young and old trees on a steep hillside (slope: 10%) east of the creek. A shallow 0.05 m litter layer characterizes the forest soils. Furthermore, there are two transects (G1, G2) located in the riparian zone at a 4 m distance to each side to the Vollnkirchener Bach. One of the two transects is managed and grazed grassland (G1), mainly covered with brown knapweed (*Centaurea jacea*), meadow foxtail (*Alopecurus pratensis*), red clover (*Trifolium pratense*) and ribwort plantain (*Plantago lanceolata*). The second transect (G2) represents a wetland and is mainly covered by meadowsweet (*Filipendula ulmaria*), common nettle (*Urtica dioica*), hoary ragwort (*Senecio erucifolius*) and field bindweed (*Convolvulus arvensis*). The groundwater table is

close to the surface on both grassland sites. The mean annual wet depositions of nitrate and ammonium were measured from 2013–2015 with 1.66 kg N ha^{-1} and 3.45 kg N ha^{-1}, respectively. In the catchment, the mean annual precipitation is 588 mm, and the mean annual temperature is 10.5 °C for the hydrological year 1st Nov. 2013 - 31st Oct. 2014 (Seifert et al., 2016). The soil moisture is measured at A3 [0.2, 0.4 and 0.6 m], at G2 [0.1 and 0.25 m] and at W1 [0.15 and 0.25 m] and has been recorded at an hourly resolution since 2013. The weekly trace gas measurements began in November 2013 and range so far until December 2015. GHG exchange fluxes were measured manually with non-steady state opaque chambers, each covering a basal area of 0.12 m^2. Chambers were placed on frames (both polypropylene), which were inserted approx. 8 cm into the soil in order to facilitate gas-tight sampling as well as to avoid soil structural damage and lateral trace gas leakage. Each chamber is equipped with an extraction septum, a counterbalance valve (in-box pressure balance) and a small fan/ventilator for homogenous mixing of the headspace air. During a 40-minute closure period, five air samples are taken from the chamber headspace at regular time intervals t0-t4 of ten minutes (0, 10, 20, 30 and 40 min.). Samples are analysed by gas chromatography (GC 8610C, SRI Instruments, Torrance, US) with an ECD for N_2O and a methanizer and FID for CO_2. Sampling was performed on a weekly basis, with five replicated chambers per transect sampled by the gas sample pooling technique (Arias-Navarro et al., 2013). According to this approach, at any time interval (t0-t4), 10 ml headspace samples are collected subsequently from any of the five replicated chambers and are pooled into one gas-tight glass vial (SRI Instruments). The trace gas fluxes are calculated from the rate of change in the headspace gas concentration over time by linear regression and were corrected for the chamber temperature, atmospheric pressure and chamber volume according to Barton et al. (2008). All measurements with a regression quality of $r^2 < 0.7$ for CO_2 (using at least four individual samples) were rejected.

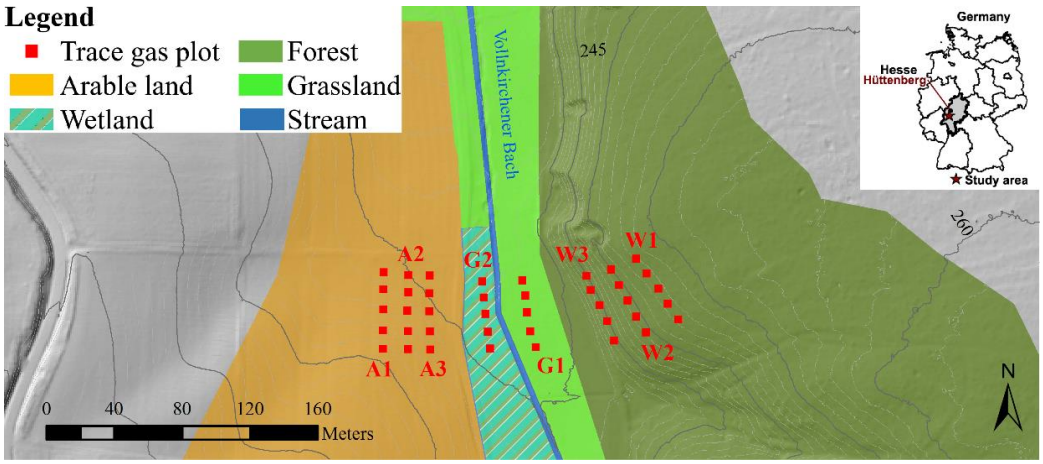

Figure III.1. Map of the study area. Red squares represent GHG chamber positions at the different transects. Dark grey contour lines represent 5 m differences in elevation, light grey areas are outside of the catchment area.

Soil emissions of CO_2 and N_2O can be subject to significant diurnal patterns, with peak values observed in the early afternoon (Savage et al., 2014), impeding the up-scaling of hourly measured emissions (usually obtained at midday) to daily values. We performed multiple linear regression (ordinary least squares regression including air temperature, relative humidity and water filled pores space) to account for the difference between, e.g., daytime (Wohlfahrt et al., 2005a) and night-time respiration (Wohlfahrt et al., 2005b). In our dataset, only CO_2 emissions showed significant correlations with the mentioned environmental drivers on arable land ($r^2 = 0.53$), grassland ($r^2 = 0.59$) and forest ($r^2 = 0.51$). Following Subke et al. (2003), we derived an hourly integration formula in order to obtain daily representative mean values of CO_2 emissions from our field measurements conducted mostly between 9 am and 5 pm. N_2O emissions are up-scaled to daily mean values with the common approach, i.e., by multiplying hourly emissions by 24. Annual CO_2 and N_2O emissions are calculated by linear interpolation between the measurements. All the underlying data is available upon request from a database (http://fb09-pasig.umwelt.uni-giessen.de:8081/).

Modelling approach

Model set up

We tested the biogeochemical model framework LandscapeDNDC (Haas et al., 2013) with the observed data from our study area. Individual models were set up for arable land, grassland and forest ecosystems. The models describe different processes in ecosystem compartments, i.e., mathematical descriptions of microclimate, water cycle, plant physiology and soil biogeochemical processes. We applied the biogeochemical model $MeTr^x$ (Kraus et al., 2015) and the water cycle model watercycleDNDC (Kiese et al., 2011) for all land uses. The biogeochemical model $MeTr^x$ simulates the turnover of soil organic matter and plant debris depending on their chemical structures (e.g., lignin and cellulose content, C/N ratio), soil properties (e.g., pH value) and meteorological drivers. Following the ‘anaerobic balloon’ concept of Li et al. (2000), major metabolites (e.g., NO_3) are distinguished between aerobic and anaerobic counterparts in order to simulate the share of nitrification and denitrification and the related production of GHG emissions. Simulated model outputs are, among others, emissions of CO_2 and N_2O. The watercycleDNDC model simulates soil water dynamics, i.e., potential evapotranspiration based on Thornthwaite and Mather (1957), transpiration depending on gross primary productivity, the water use efficiency of the modelled plant types and soil water flow based on a cascading bucket model approach (Kiese et al., 2011). The latter determines the advective transport of nutrients into deeper soil layers. model driving data, i.e., meteorological data and land use-specific soil and vegetation characteristics. To simulate plant growth on the three different land use types, we selected the individual physiology modules

arableDNDC, grasslandDNDC (Kim et al., 2015; Molina-Herrera et al., 2016) and PSIM (Grote et al., 2009).

Table III.1. Input settings of the LandscapeDNDC model for the three different land uses in the Vollnkirchener study region, based on measurements and farmers management documentation. In case spans are given, they reflect observed ranges for measurements used throughout the set up of the soil profile, given from the top layer setting to the bottom layer. The soil depth was estimated for model set up. F = fertilizer application, M = manure application.

Input	Arable (A1-3)				Grassland (G1)		Forest (W1-3)	Unit
Vegetation type	Sep 10 - Jul 11	Winter Barley			Perennial grass		Light beech forest	-
	Aug 11 - Aug 12	Rape						
	Oct 12 - Aug 13	Winter Wheat						
	Oct 13 - Aug 14	Triticale						
	Sep 14 - Aug 15	Triticale						
	Oct 15 - Jul 16	Rape						
Management	2 Mar12	166.5	kg N ha^{-1}	F	01 Feb 13	Grazing		-
	2 Apr 12	49.9	kg N ha^{-1}	F	01 May13	Harvest		
	8 Nov 12	56.2	kg N ha^{-1}	F	01 Sep 13	Grazing		
	11 Mar 13	54.0	kg N ha^{-1}	F	02 Mar 14	Grazing		
	23 Apr 13	53.8	kg N ha^{-1}	F	01 May 14	Harvest		
	3 May 13	29.3	kg N ha^{-1}	M	01 Sep 14	Grazing		
	3 May 13	538.0	kg C ha^{-1}	M	20 Jan 15	Grazing		
	12 Nov 13	29.0	kg N ha^{-1}	M	29 Jun 15	Harvest		
	12 Nov 13	533.0	kg C ha^{-1}	M	26 Sep 15	Grazing		
	11 Mar 14	54.0	kg N ha^{-1}	F				
	1 Apr 14	53.8	kg N ha^{-1}	F				
	8 May 14	40.5	kg N ha^{-1}	F				
	22 Sep 14	149.0	kg C ha^{-1}	M				
	22 Sep 14	8.1	kg N ha^{-1}	M				
	8 Nov 14	1032	kg C ha^{-1}	M				
	8 Nov 14	56.2	kg N ha^{-1}	M				
	11 Mar 15	1564	kg C ha^{-1}	M				
	11 Mar 15	85.1	kg N ha^{-1}	M				
	10 Apr 15	59.4	kg N ha^{-1}	F				
	30 Aug 15	59.4	kg N ha^{-1}	F				
	12 Nov 15	29.0	kg N ha^{-1}	M				
	12 Nov 15	532.0	kg C ha^{-1}	M				
Soil texture	Sandy clay loam				Sandy clay loam		Sandy clay loam	-
Soil type	Stagnic Luvisol				Gleysol		Cambisol	
Bulk density	1.55–1.60				1.20–1.44		1.36–1.49	g cm^{-3}
Organic carbon	1.57–0.91				2.55–0.71		3.61–1.73	%
Total soil nitrogen	0.16–0.09				0.29–0.08		0.21–0.11	%
Clay content	23–26				24–25		24–26	%
pH	6.45				4.42		3.5–5.5	-
Soil depth	2.00				0.50		0.55	m

All models refer to a one-dimensional soil column, i.e., assuming homogeneous conditions in lateral directions, and were run with a daily time step resolution. Tab. 1 provides an overview of the major Arable soils are stagnic luvisols with a thick loess layer, modelled down to 2.0 m with 80 layers, while the actual soil depth is unknown. Gleysols in the meadow grassland site were modelled down to 0.5 m (set up with 40 layers), corresponding to the mean annual groundwater table depth. The thin and stony soil at the forest site is a cambisol and modelled down to bedrock (0.55 m, set up with 45 layers) with a litter height of 0.05 m. The bulk density increases with depth for every land use, while soil organic carbon and nitrogen decrease with depth. We run simulations for all land uses at a daily time resolution for 6 years, starting on 1st January 2010, using the data from Table III.1 as initialization and using a model spin-up time of two years.

Model-data fusion

For the multi-objective Bayesian model calibration, we used a two-tiered Generalized Likelihood Uncertainty Estimation (GLUE) approach (Beven and Binley, 1992). The model was iterated in both tiers 100,000 times by changing the parameter sets using Latin hypercube sampling with the Python software SPOTPY (Houska et al., 2015). The parameters for the physiology and the water-cycle modules were treated as land use-specific, while the parameters of the biogeochemical model were calibrated using the data from all land uses (Table A1). We presuppose no prior knowledge besides the given parameter ranges, i.e., we assume a uniform (non-informative) prior probability distribution for all parameters. We statistically judged the performance of every parameter set to reproduce measurements with a root mean squared error (RMSE). Similar to Bloom and Williams (2015), we do not explicitly consider measurement uncertainty during the model data fusion. As shown in Houska et al. (2017), one-tier GLUE based multi-objective model calibration can result in very low acceptance rates, down to 0.01%. We therefore considered a two-tier GLUE approach in order to increase the identifiability and accuracy of the accepted model runs:

Tier I: In the first step, we constrained the parameter space of the hydrology and plant physiology modules of LandscapeDNDC by investigating the respective parameters of both models (Table A1). We accepted only model runs that were within the best 5% of all simulated RMSEs in terms of the respective variable (WFPS at different depths [arable land at 0.2, 0.4 and 0.6 m, grassland at 0.1 and 0.25 m and forest at 0.15 and 0.25 m], as well as yield on arable land). Parameter sets were accepted if they belonged to the 5% best model runs for each land use. That is, we took the best 5% of the RMSEs for each respective output variable and took only the intersecting parameter set, which are all from the selected variables for one land use. The results of tier I are summarized in supplementary Fig. A1-A4 and are not further discussed in this study, as they belong to the initialization of the model.

Tier II: To achieve realistic GHG simulations from the MeTr^x biogeochemical module of LandscapeDNDC, we took the posterior parameter boundaries of tier I and ran GLUE with all parameters of Table A1 again. This time, we considered the best 5% of all RMSEs in terms of the respective N_2O and CO_2 emissions for each land use (A1-3, G1 and W1-3). Again, only the 5% best intersecting parameter sets were accepted per land use. These results are shown in the following chapters. There was no major effect of the biogeochemical model parameters on the WFPS simulation.

Posterior model runs of tier II were further investigated in three different ways:
(1) Seasonal comparisons of measured and modelled emissions for spring (21st March - 20th June), summer (21st June - 20th September), autumn (21st September - 20th December), and winter (21st December - 20th March).
(2) Management comparison of measured and modelled emissions, i.e., investigation of model performance within two weeks before and two weeks after management events to check model performance in generating hot moments, e.g., after fertilizer application.
(3) Model performance in simulating magnitude and uncertainty of C and N fluxes not measured *in situ*, such as N_2 or autotrophic and heterotrophic components of CO_2 emissions.

Results and discussion

Measured N_2O fluxes

To determine the representativeness of each transect for a given land use, the respective differences in measured N_2O emissions were compared (Table III.2). The temporal dynamics of N_2O emissions are presented (Figure III.2), distinguishing between different seasons (Figure III.3) and before/after management events (Figure III.4).
Arable land N_2O fluxes: Emissions on arable land vary between 0 and 0.3 kg N_2O-N ha^{-1} day^{-1}. There were no significant differences over time between the three weekly measured transects on arable land (Table III.2). The highest emissions occur mostly after management events. Mineral fertilizer application in particular stimulates N_2O emissions, causing hot moments from, for example, March to May 2014. The input of N through manure application has a minor influence on the magnitude of N_2O emissions. The mean annual measured N_2O emissions from arable land are comparably high with 4.5 kg N_2O-N ha^{-1} a^{-1} (Jungkunst et al., 2006), equalling a GWP of 575 kg CO_2-C equiv. ha^{-1} a^{-1}. With a yearly fertilizer application of 248.2 kg N a^{-1} a mean annual emission factor (EF) of 1.4% (varying between 1.2% for A2 and 1.8% for A3) can be calculated, where 1 kg N ha^{-1} a^{-1} is attributed to the background emissions of unfertilized soil (IPCC, 1997). This EF is inside the IPCC-assumed range of 1.25 ±1% and close to the average EF (1.56%) of several (n=56) agricultural sites in

Germany (Jungkunst et al., 2006). A robust finding throughout the literature is that reduced nitrogen input would lead to lower emissions and therefore more climate-friendly agriculture (Bouwman et al., 2002).

Grassland N_2O fluxes: N_2O emissions significantly vary between the grazed site G1 and the wetland site G2, which can be attributed to differences in management, hydrological, soil and vegetation characteristics. Most likely, the nitrate supply through groundwater and uptake by the rooting system of the plants is important (Liebermann et al., 2017). Even though the groundwater table (0.2 - 0.4 m belowground) is rather shallow in the winter/spring, the uptake rates in summer/autumn (groundwater table 0.3 - 1.0 m belowground) are supposedly larger due to the vegetation period. Here, capillary rise may play a relevant role (Orlowski et al., 2016). G1 is characterized by a mix of *Centaurea jacea*, *Alopecurus pratensis*, *Plantago lanceolata* and *Trifolium pratense*, is grazed by sheep twice a year and is cut once a year. Emissions from the grazed grassland vary between -0.0019 and 0.014 kg N ha^{-1} day^{-1}. High emissions were measured after grazing, e.g., in March 2014 when sheep dung was stimulating N_2O emissions. Negative values depict N_2O uptake and are frequently found under prevailing wet conditions in spring, a finding that was also reported by Glatzel and Stahr (2001). The grassland annual N_2O emissions are much lower than those observed for the arable system (A1-3). However, with 0.29 kg N_2O-N ha^{-1} a^{-1} are they in accordance with a study site 12 km northeast of our site, where annual emissions range from 0.18 to 0.79 kg N_2O-N ha^{-1} a^{-1} on an unfertilized grassland with shallow groundwater table (Kammann et al., 1998). Their study also reports a similar seasonal pattern to our measurements, with emissions close to zero in the dry and colder autumn months. The measured annual emissions are below the assumed background level of N_2O-N emissions of 1 kg N_2O-N ha^{-1} a^{-1} from agricultural soils (IPCC, 1997). The annual N_2O emissions are equal to a GWP of 37 kg CO_2-C equiv. ha^{-1} a^{-1}. The EF through grazing is 3.8%, which is in accordance with typical emissions factors from extensive grazed grasslands, ranging globally from 0.2 - 9.9% (Oenema et al., 1997).

Wetland N_2O fluxes: The non-managed transect G2 is dominated by species such as *Urtica dioica*, *Filipendula ulmaria* and *Senecio erucifolius.* Typically, a deeper rooting system is found compared to that in the grazed grassland transect G1, and accordingly, additional nitrate uptake from the groundwater is more prevalent. The mean measured emissions are higher on the non-managed G2 than on the grazed G1 throughout the year, especially during summer and autumn (Figure III.3). The annual emissions are accordingly nearly two times higher at 0.52 kg N_2O-N ha^{-1} a^{-1}, which is equal to a GWP of 66 kg CO_2-C equiv. ha^{-1} a^{-1}.

Forest N_2O fluxes: Significant differences were found for the forest transects W2 and W3, which can be explained by natural variations along the steep hillslope: On the hillside (W2) the soil is potentially

washed out through lateral transport, leading to decreased nutrient availability, compared to the drier top (W1, +200% N_2O emissions) and the wetter hillfoot (W3, +330% N_2O emissions). The N_2O emissions from the forest transects are mostly low, ranging between -0.003 and 0.004 kg N ha^{-1} day^{-1}. Higher emissions were measured only for several weeks in January 2014, with the highest values observed at W1. We attribute this to freeze-thaw effects, typically found when year-around measurements are considered (Papen and Butterbach-Bahl, 1999). Negative fluxes were measured, for example in March and May 2014. The underlying process of N_2O uptake has been reported before (e.g., Flechard et al., 2005; Neftel et al., 2007) and is assumed to be a microbial process, in which denitrifiers use N_2O as an electron acceptor for respiration under wet/anaerobic conditions (Bremner, 1997). Negative emissions occur during times with high WFPS (Fig. A3), which is in accordance with Bremner (1997). However, our measured negative emissions are low compared to the variance between transects (W1-3), i.e., they could also originate from measurement errors. Our annual measured emissions in forests are 0.08 kg N_2O-N ha^{-1} a^{-1} (GWP of 10 kg CO_2-C equiv. ha^{-1} a^{-1} CO_2 emissions), which is much lower than that at adjacent grassland and arable sites. Moreover, this value is almost two orders of magnitude lower than the N_2O emissions (5.1 kg N_2O-N ha^{-1} a^{-1}) measured from a beech forest in Högelwald, Germany (Papen and Butterbach-Bahl, 1999). A likely reason is the substantially higher annual deposition rate of 25 kg N ha^{-1} a^{-1}, an N input five times higher than that in our system. However, our measurements of N deposition only include wet deposition. Additional dry depositions are often assumed to add another 30-60% to total atmospheric N deposition (Flechard et al., 2011).

Table III.2. Mean measured annual fluxes (Nov 2013 - Dec 2015) on the different land use transects of the Vollnkirchener Bach study area. Differences between the investigated transects and land uses for measured and modelled N_2O emissions in kg N-N_2O ha^{-1} a^{-1}. * = significant difference ($p < 0.05$, Kruskal-Wallis test). Arable (A1-3), Grassland (G1), Wetland (G2), Forest (W1-3), RMSE in kg N-N_2O ha^{-1} day^{-1}.

	A1	A2	A3	G1	G2	W1	W2	Measured	Mean measured	Mean simulated	Posterior RMSE
A1								4.08			0.0326 - 0.0353
A2								3.87	4.49	7.33	0.0238 - 0.0278
A3								5.53			0.0285 - 0.0329
G1	*	*	*					0.29	0.29	0.69	0.0029 - 0.0038
G2	*	*	*	*				0.52	0.52	-	not simulated
W1	*	*	*		*			0.09			0.0022 - 0.0025
W2	*	*	*	*	*			0.03	0.08	0.33	0.0014 - 0.0021
W3	*	*	*		*		*	0.13			0.0018 - 0.0021

Measured CO_2 fluxes

Emissions measured using our closed chamber on arable land and grassland include those from soil and vegetation, as entire plants are covered by the chamber. Therefore, we interpret these emissions as total ecosystem respiration (TER). In contrast, chambers in the forest were placed on the forest floor without any vegetation inside; thus, these measurements include soil (heterotrophic) and root

(autotrophic) respiration, i.e., below ground respiration only. To determine the representativeness of each transect for a given land use, the respective differences in measured CO_2 emissions were compared to each other (Table III.3). The measured CO_2 emissions are given over time (Figure III.5), separated into different seasons (Figure III.6) and before/after management-events occur (Figure III.7).

Arable TER: Measured values from our arable transects range between 0 to 175.2, 199.6 and 143.1 kg C-CO_2 ha^{-1} day^{-1} for A1, A2 and A3 respectively and are not significantly different between the transects (compare Table III.3). Emissions occur mainly during the growing season, starting in March and ending in November. For a comparable study site in southern Finland, reported daily TER values under barley were between 23.6 to 235.6 kg C-CO_2 ha^{-1} day^{-1} during May and September (Lohila et al., 2003), which is in the same range as our observations. The annual sum of our TER emissions is 19.96 ± 2.36 t C-CO_2 ha^{-1} a^{-1}. This is slightly lower than yearly TER measured on a winter wheat study site in Belgium with 23.18 t C-CO_2 ha^{-1} a^{-1} (Suleau et al., 2011). Demyan et al. (2016) reported lower values, with an average total of 11.43 t C-CO_2 ha^{-1} a^{-1}, derived from observations spanning six growing seasons in southwestern Germany. However, all studies are possibly prone to overestimations of the emissions from September to November, as daily emissions are generated with a multiple linear regression model, and in our case, are based on our hourly measurements of air temperature and soil moisture. Such methods do not fully account for management effects, such as harvests (Subke et al., 2003).

Grassland TER: Emissions from grassland vary from 5.0 to 68.3 t C-CO_2 ha^{-1} a^{-1}, with no significant difference between the two transects G1 and G2. Emissions are close to zero in the winter months (December to February) and highest during the growing season. A distinct negative correlation between the measured TER with WFPS was found during wet conditions from end of June to July in 2014. In this time, emissions decrease to 41.0 kg C-CO_2 ha^{-1} day^{-1}. The total yearly emissions are 11.79 t C-CO_2 ha^{-1} a^{-1}, which agrees well with the mean yearly emissions reported for 19 different grassland sites across Europe, with mean annual emissions of 12.83 t C-CO_2 ha^{-1} a^{-1} (Gilmanov et al., 2007). However, due to the many different grassland sites considered in their study, Gilmanov et al. report a much wider range of observed annual TER values, from 4.9 to 16.4 t C-CO_2 ha^{-1} a^{-1}. They also found that management is a main influencer of TER, where intensively managed grasslands produce higher emissions than extensively managed grasslands. With regard to grazing, we found only a minor direct impact on the measured flux rates (Figure III.7).

Wetland TER: Emissions from the study site G2 vary from 0 to 92 kg C-CO_2 ha^{-1} day^{-1} and are higher than those from G1, especially in the growing season. This is due to the higher above ground biomass of the different species present and represents a common pattern in unmanaged grasslands (Soussana

et al., 2007). Emissions typically end with the cessation of pasture growth during temperatures under 5°C (Parsons, 1988). The annual emissions are 12.54 t C-CO_2 ha^{-1} a^{-1}, driven by the growing season. Forest below ground respiration: The mean measured belowground respiration spans between minimum values of 2.1 to 4.5 and maximum values of 9.3 to 19.9 kg C-CO_2 ha^{-1} day^{-1} between the different transects (W1-3). While we found higher emissions in the summer months, seasonal differences have a lower magnitude of TER on arable and grassland. This was expected, as we do not measure above ground biomass respiration on our forest study site. Overall, rewetting has the strongest influence on changes in belowground respiration in our forest study sites. The highest emissions occurred in July 2014 after several rewetting events of the uppermost soil layer (Fig. A1). Xiang et al. (2008) reported that multiple rewetting leads to respiration rates of up to eight-times higher. The total yearly soil emissions are 2.98 ± 0.89 t C-CO_2 ha^{-1} a^{-1}, which is at the lower end of other European forest ecosystems, e.g., 6.6 ± 2.9 t C-CO_2 ha^{-1} a^{-1}, as reported by Janssens et al., (2001). The uphill transect W1 has the highest emission rates throughout the year and shows significant differences when compared to W2 and W3. This transect is less shaded by trees, resulting in a 1.3°C higher annual mean soil temperature compared to W2 and W3, likely causing higher CO_2-emissions (Table III.3).

Table III.3. Mean measured annual fluxes (Nov 2013 - Dec 2015) from the different land use transects of the Vollnkirchener Bach study area. Differences between the investigated transects and land uses for measured and modelled CO_2 emissions in t C-CO_2 ha^{-1} a^{-1}. * = significant difference ($p < 0.05$, Kruskal-Wallis test). Arable (A1-3), Grassland (G1), Wetland (G2), Forest (W1-3), RMSE in kg C-CO_2 ha^{-1} day^{-1}.

	A1	A2	A3	G1	G2	W1	W2	Measured	Mean measured	Mean simulated	Posterior RMSE
A1								20.10			30.73 - 36.38
A2								22.25	19.96	20.53	35.66 - 42.26
A3								17.54			22.90 - 28.46
G1								11.79	11.79	13.24	7.01 - 9.08
G2								12.54	12.54	-	not simulated
W1	*	*	*	*	*			4.00			3.53 - 3.89
W2	*	*	*	*	*	*		2.38	2.98	3.28	3.37 - 4.07
W3	*	*	*	*	*	*		2.56			3.15 - 3.96

Modeled N fluxes

After selecting the posterior model runs as described in the model-data fusion chapter, we found the model to be generally capable of reproducing the measured data and consequently investigated the modelled C and N cycles in more detail. The modelled N_2O emissions are shown for the different land uses over time (Figure III.2), separated into different seasons (Figure III.3) and before/after management-events occur (Figure III.4). The complete modelled N cycle is given in Table III.4.

Arable land N cycle: The arable land simulations consider an annual N input of 198 kg N ha^{-1} a^{-1}. This input is balanced by 108.6 ± 50.1 kg N ha^{-1} a^{-1} gaseous (primarily N_2), 30.0 ± 29.9 kg N ha^{-1} a^{-1} nitrate leaching and 99.7 ± 7.8 kg N ha^{-1} a^{-1} harvest losses (Table III.4), meaning that the modelled

outputs are higher than the given inputs. This gap in the annual N cycle is fed by soil storage in the model, indicating N depletion over time. Even though N losses through NO_3^- and particularly N_2O emissions (7.3 ± 2.3 kg N ha^{-1} a^{-1}) are only a minor proportion of the total N balance, both rates are high regarding their environmental impacts as a GHG contributing to global warming and as a water pollutant regarding eutrophication and drinking water supply, respectively. However, the uncertainty related to our estimated NO_3^- leaching rate is overall the largest source of uncertainty in our N balance. These estimates cannot be sufficiently constrained with the given observation data, but they are in accordance with other reported N leaching rates on arable land in Germany (Siemens and Kaupenjohann, 2002).

The simulated N_2O emissions contribute 3.1% to the total simulated N losses. The underlying model runs follow the trend of the observation data. Hot moments can be observed after fertilizer applications, and they are predicted by the model in time but sometimes not in magnitude (e.g., March to May 2014). During these events, soil moisture is often not modelled accurately: The model predicts rewetting processes that have not been measured at the same magnitude (Fig. A1), which might explain the overestimated fluxes. One possible reason may also be uncertain rainfall model input data. Kavetski et al. (2006) found the measurements of precipitation within a catchment to be uncertain, as the trajectory of storm cells through a catchment may be different for each storm and may not have their centres at the rain gauge, where rainfall inputs are traditionally measured. Our rainfall data are measured 4 km northeast of the trace gas study area and is likely affected by such uncertainties.

The total simulated and measured emissions on the arable site are highest in the spring (Figure III.3). While the transects A1 and A2 vary, with 95% of the values between 0 and 0.05 kg N_2O-N ha^{-1} day^{-1}, A3 shows more variation, up to 0.15 kg N_2O-N ha^{-1} day^{-1}. As A3 is located at the hill toe, we attribute this effect to the lateral transport of nitrate from uphill. However, our one-dimensional model setup does not cover lateral water and nutrient transport; accordingly, the model is not able to predict the higher emissions at A3 in the spring. While such a process is part of complex integrated hydro-biogeochemical catchment models (Haas et al. 2013; Klatt et al., 2017; Wlotzka et al., 2013), it has not yet been confirmed experimentally. The distributions of the measured emissions in the summer, autumn and winter seasons are well in accordance with the modelled emissions. Furthermore, the modelled emissions are also in agreement with emissions measured before and after manure applications (Figure III.4). This result agrees with a study by Molina-Herrera et al., (2016) who found LandscapeDNDC to be capable of simulating agricultural N_2O emissions. However, in our case, the model overestimates peak emissions before fertilizer applications, which leads to higher mean annual modelled emissions (7.33 kg N_2O-N ha^{-1} a^{-1}). This is 2.8 kg N_2O-N ha^{-1} a^{-1} higher than our observed emissions and is even outside the large model uncertainty of 2.3 kg N_2O-N ha^{-1} a^{-1}. Hence, future

research should specifically investigate the reason for this overestimation of peaks, either by revising the model structure or by identifying other sources of model uncertainty.

Grassland N cycle: Grassland simulations consider an annual N input of 12.7 kg, with 7.6 kg coming from modelled biomass that is transferred into dung and urine applied by grazing sheep. The simulated N loss is substantially larger than the N input, with 22.3 ± 13.3 kg N ha^{-1} a^{-1} gaseous losses (primarily N_2), 1.5 ± 3.19 kg N ha^{-1} a^{-1} occurring as nitrate leaching and 29.8 ± 9.4 kg N ha^{-1} a^{-1} as biomass removal through grazing sheep and harvest (hay making). Comparing inputs and outputs, we simulated a mean nitrogen gap of 40.9 ± 25.9 kg N ha^{-1} a^{-1}. The model suggests decreasing soil organic N stocks. So far, we have only initial measurements of soil organic N content. However, we assume that the source of additional N in the form of nitrate in shallow groundwater is a potential dominating process that is not included in the current LandscapeDNDC version we used. Liebermann et al. (2017) used a revised LandscapeDNDC setup for hypothesis testing to identify potential additional N sources in groundwater-dominated grasslands and showed that groundwater N uptake is a likely contributor.

Taking a closer look at the modelled N_2O emissions, one can see that the model did not reproduce high or negative (N_2O uptake) emissions. Currently, LandscapeDNDC does not consider any N_2O uptake, and accordingly, negative fluxes cannot be simulated by the model. The peaky dynamics of the simulated N_2O emissions, especially from August 2014 to January 2015, are not confirmed by the measurements, indicating possible measurement errors during this period of time. In a grazed system with, in our case, approximately 70 sheep per hectare, the animal urine patches create emissions hot spots. With only five chambers, it is possible that the measurements could miss these hot spots. Additionally, the LandscapeDNDC model will assume that the manure is uniformly spread over the field, producing emissions that are likely to be higher than those from non-urine patches, but lower than those from urine patches. One has also to consider the temporal mismatch of our weekly N_2O measurements and the hourly simulations, making a full match of the observations with the simulations difficult. So far, there is no clear effect of grazing on the N_2O emissions on the grassland site in both the measurements and modelled results (Figure III.4). The mean modelled annual emissions overestimate the observations by 0.4 kg N_2O-N ha^{-1} a^{-1}, and even the simulated uncertainty bounds of 0.27 kg N_2O-N ha^{-1} a^{-1} do not capture the measured dynamics.

Forest N cycle: The N input is given for the forest model only considering atmospheric deposition with an annual amount of 5.1 kg N ha^{-1} a^{-1}. Gaseous losses amount to 1.8 ± 2.0 kg N ha^{-1} a^{-1}. Leaching contributes to 2.0% of the N output. The rest (3.3 ± 2.0 kg N ha^{-1} a^{-1}) is allocated into biomass and soil. By taking a closer look at the N_2O emissions (Figure III.2), we see that the model fails to reproduce the observed emission dynamics.

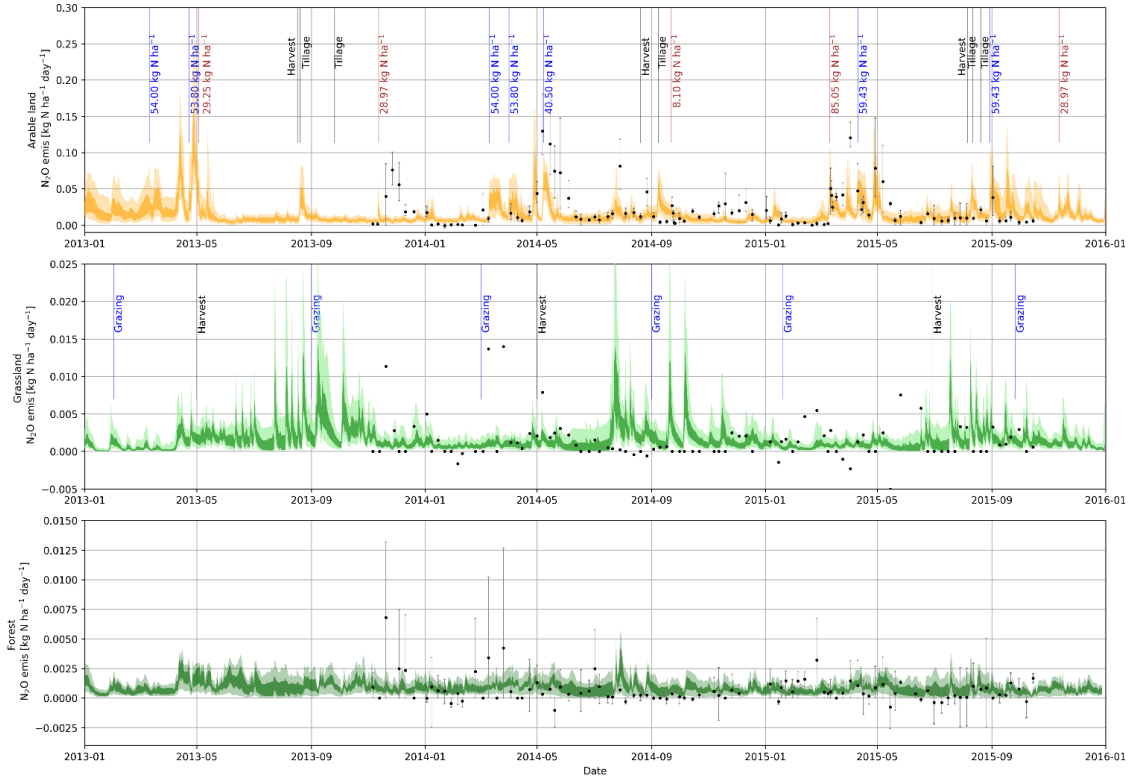

Figure III.2. Measured and modelled N_2O emissions from different land use. Measurements are given as grey error bars showing the variance between the replicated transects and the mean value as a black dot. Posterior model uncertainty is given in light colour for the 5 and 95 percentiles and dark colour for the 25 and 75 percentiles. Vertical lines indicate management events. In the uppermost panel, blue coloured vertical bars indicate fertilizer application, while brown colours indicate manure application.

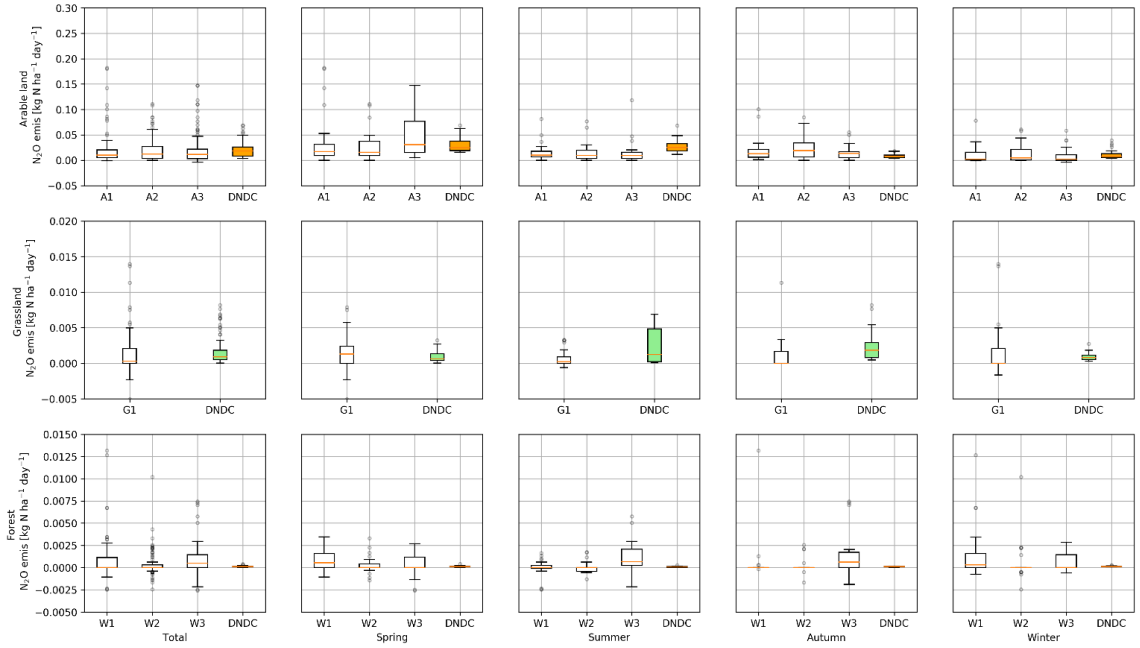

Figure III.3. Observed and modelled N_2O emissions for spring (21st Mar. - 20th Jun.), summer (21st Jun. - 20th Sep.), autumn (21st Sep. - 20th Dec.), and winter (21st Dec. - 20th Mar.).

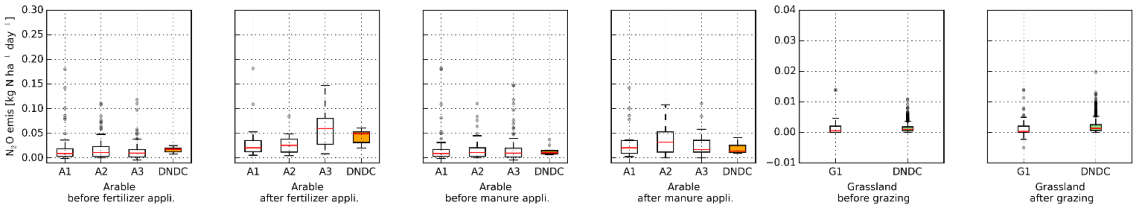

Figure III.4. Management effects on N_2O emissions. Measured and modelled emissions within a time window of 2 weeks before and 2 weeks after a management.

Table III.4. Simulated nitrogen fluxes given by posterior model runs and their uncertainty on different land use in [kg N ha^{-1} a^{-1}]. N manure on grassland includes urine and dung input by sheep. Biomass output on grasslands combines harvest export and biomass leaving the system through sheep. Arable land model assumes 20% return of stubble to field.

Modeled N flux	**Arable land**		**Grassland**		**Forest**	
N deposition	5.11		5.11		5.11	
N manure	57.55		7.57		0	
N fertilizer	135.37		0		0	
Total input	**198.03**		**12.68**		**5.11**	
NO emis.	0.57	±0.16	0.46	±0.21	0.45	±0.33
N_2 emis.	62.55	±26.83	18.69	±10.91	1	±1.5
N_2O emis.	7.33	±2.3	0.69	±0.27	0.33	±0.15
NH_3 emis.	38.15	±20.8	2.45	±1.89	<0.01	±<0.01
Total gaseous output	**108.6**	**±50.09**	**22.29**	**±13.28**	**1.78**	**±1.98**
DON leaching	0.01	±<0.01	0.01	±<0.01	0.01	±<0.01
NO_3 leaching	30.01	±29.9	1.46	±3.19	0.03	±0.04
Total leaching output	**30.02**	**±29.9**	**1.47**	**±3.19**	**0.04**	**±0.04**
N grain export	63.92	±5.17	0		0	
N straw export	35.75	±2.67	29.77	±9.44	0	
Total biomass output	**99.67**	**±7.84**	**29.77**	**±9.44**	**0**	
Balance	**-40.26**	**±87.83**	**-40.85**	**±25.91**	**3.29**	**±2.02**

The observed N_2O emissions have high error bars, and not all transects are driven by frost-thaw cycles or N_2O uptake at the same time (Table III.2). Parameterizing and simulating the forest transects independently from each other would improve the simulations. One limiting factor is that both N_2O uptake and frost-thaw cycles are not included in the current version of LandscapeDNDC. We therefore recommend the inclusion of frost-thaw cycles (e.g., based on De Bruijn et al., 2009) in the model, as this process can have a major influence on N_2O inventories, e.g., up to 73% of the total annual N_2O loss at a forest site in Högelwald, Germany (Papen and Butterbach-Bahl, 1999). The mean modelled annual emissions (0.33 ± 0.15 kg N ha^{-1} a^{-1}) overestimate the observed emissions on all transects.

Modeled C fluxes

The modelled CO_2 emissions are shown for the different land uses over time (Figure III.5), separated into different seasons (Figure III.6) and before/after management events (Figure III.7). The complete modelled C cycle is given in Table III.5.

Arable land C cycle: The LandscapeDNDC simulations for the arable system predict a mean annual gross carbon uptake of 25.7 ± 1.3 t C-CO_2 ha^{-1} a^{-1}. 20.5 ± 1.8 t C-CO_2 ha^{-1} a^{-1} leaves the system through respiration, to which maintenance respiration contributes the largest proportion (65%). This is in accordance with annual measured losses (Table III.3). The harvest output is with 4.7 ± 0.4 t C ha^{-1} a^{-1} and is in good agreement with the observed yields (Figure A4). However, the temporal dynamics of the modelled TER on the arable land study site underestimate the emissions in the summer season (Figure III.6), and the mean modelled fluxes are substantially lower than those measured before and after the harvest (Figure III.7). Tillage and harvest events occur in the summer season. While the observed emissions drop after harvest by 25%, the modelled emissions drop by 50%. The reason for this is either an underestimation of the emissions through LandscapeDNDC (after harvest events until tillage occurs) or uncertainties in the measured CO_2 emissions upscaling method (discussed in chapter 2.1). As microbial processes can oxidize more soil carbon after harvests (resulting in higher heterotrophic respiration), we assume that the discrepancy stems from the model simulations. There are studies, e.g., Buyanovsky et al. (1986), which report the highest soil respiration rates after harvests. The modelled and measured soil CO_2 emissions agree well after tillage. However, unless there is a gap of two weeks or more between harvest and tillage, the "pre-tillage" results will include some post-harvest effects, and the "post-harvest" results will also include some post-tillage effects. Our intention to present the data grouped by these events are the discrepancies between modeled and observed CO_2 dynamics. There is a sharp drop of modeled CO_2 emissions after harvest due to the prompt absence of autotrophic respiration. In reality, there will likely be some ongoing metabolic respiration of plant tissue remaining in the field, which is not represented by the 'assumed' dead plant material in the model. After incorporation of harvest residues (at tilling) modeled CO_2 emissions increase again sharply. The sharp increase is due to the incorporation and hence availability of fresh litter (stubble) and a temporary stimulation of decomposition by the model due to the disruption/aeration of the soil structure. Both, overestimation of fresh litter and/or stimulation of decomposition by the model may contribute to the discrepancies between observed and modelled CO_2 emissions.

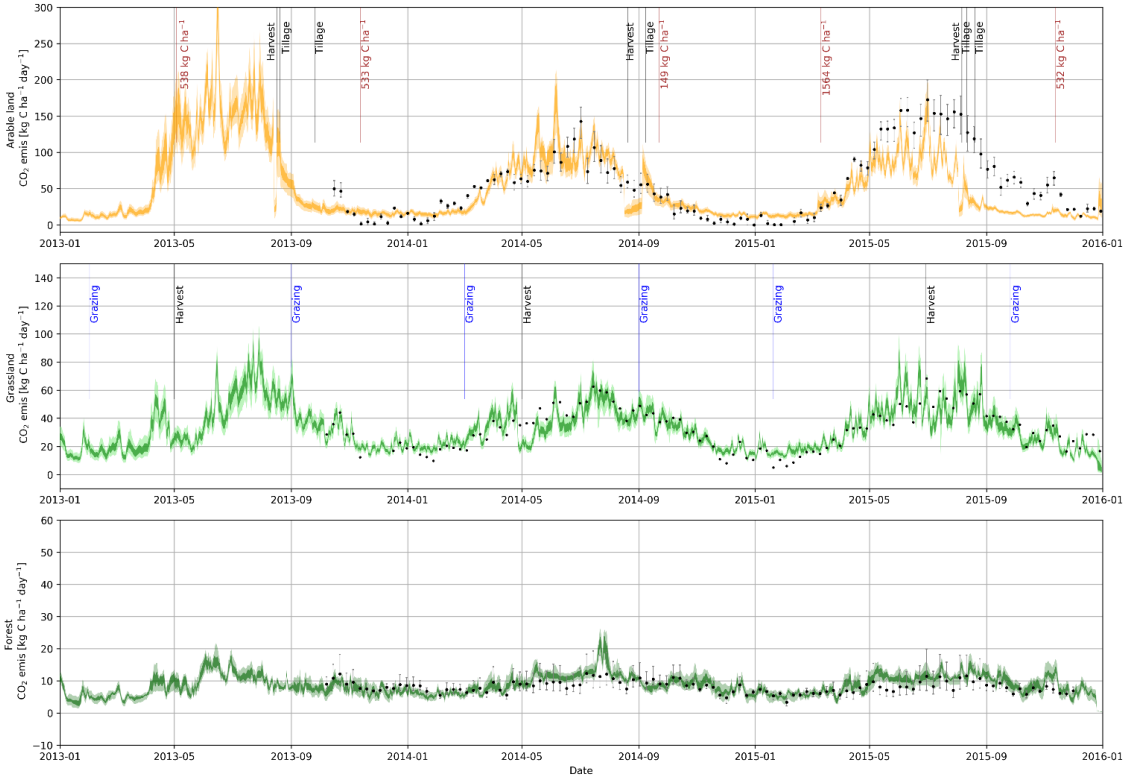

Figure III.5. Modelled CO_2 emissions and management. Measurements are given as grey error bars showing the variance between the replicated transects and the mean value as a black dot. Posterior model uncertainty is given in light colour for the 5 and 95 percentiles and dark colour for the 25 and 75 percentiles. Vertical lines indicate management events. Brown coloured bars in the uppermost panel indicate manure application.

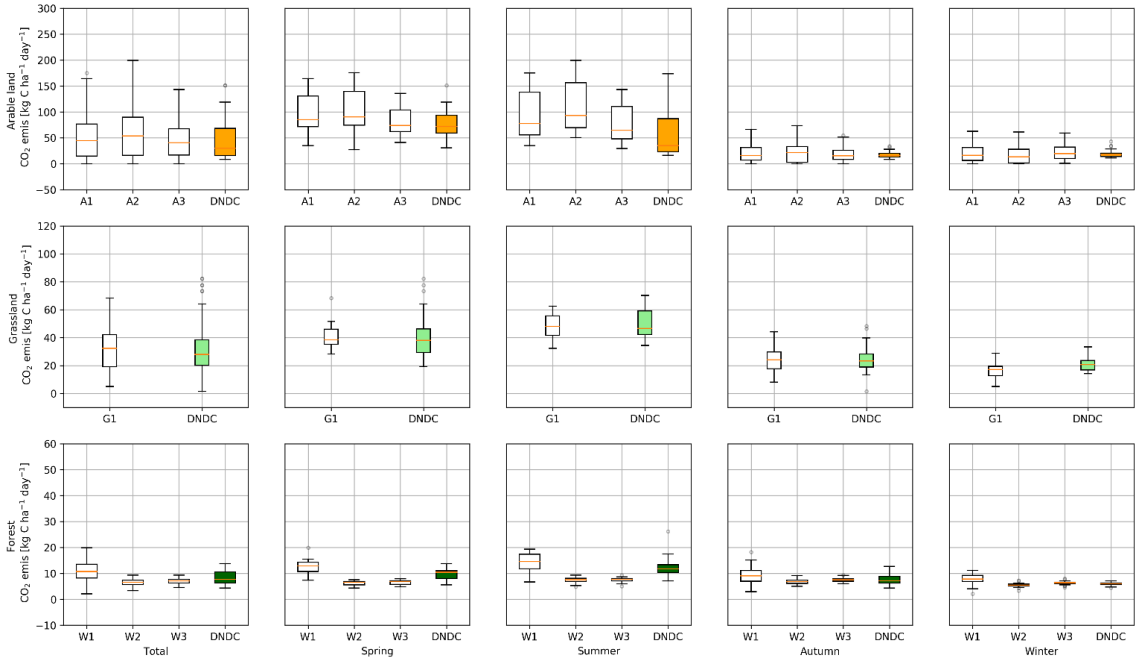

Figure III.6. Observed and modeled CO_2 emissions for spring (21.03 20.06.), summer (21.06. 20.09.), autumn (21.09. 20.12.), and winter (21.12. 20.03.).

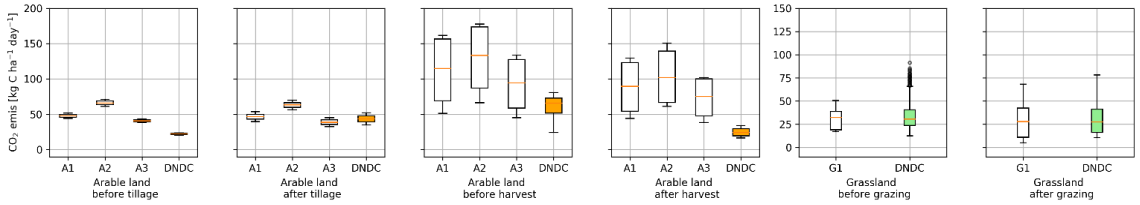

Figure III.7. Management effects on CO_2 emissions. Measured and modeled emissions where selected in a time window of 2 weeks before and 2 weeks after a management.

Table III.5. Simulated carbon fluxes given by posterior model runs and their uncertainty on different land use in [t C ha^{-1} a^{-1}]. C manure on grassland includes input by sheep's dung. Arable land model assumes 20% return of stubble to field.

Modeled C flux	**Arable land**		**Grassland**		**Forest**	
CO_2 uptake	24.65	±1.32	16.8	±1.72	8.94	±0.56
C manure	1.06		0.07		0	
Total input	**25.71**	**±1.32**	**16.87**	**±1.72**	**8.94**	**±0.56**
Growth respiration	2.53	±0.2	0.81	±0.27	1.44	±0.05
Heterotrophic respiration	4.69	±0.53	2.27	±0.9	2.04	±0.1
Maintenance respiration	13.31	±1.06	10.16	±1.13	3.11	±0.39
Total gaseous output	**20.53**	**±1.79**	**13.24**	**±2.3**	**6.59**	**±0.54**
DOC leaching	<0.01	±<0.01	<0.01	±<0.01	<0.01	±<0.01
Total leaching output	**<0.01**		**<0.01**		**<0.01**	
C bud export	1.97	±0.17	0		0	
C straw export	2.75	±0.21	2.28	±0.72	0	
Total biomass output	**4.72**	**±0.38**	**2.28**	**±0.72**	**0**	
Balance	**0.46**	**±3.49**	**1.35**	**±4.74**	**2.35**	**±1.1**

Grassland C cycle: The LandscapeDNDC simulations for the grassland system (G1) predict a mean annual gross carbon uptake of 16.9 ± 1.7 t C-CO_2 ha^{-1} a^{-1} and an annual loss of 13.2 ± 2.3 t C-CO_2 ha^{-1} a^{-1} through respiration. The rest is related to grazing (0.2 ±< 0.01 t C-CO_2 ha^{-1} a^{-1}), harvesting (2.1 ± 0.7 t C-CO_2 ha^{-1} a^{-1}) and allocation in the soil (1.4 ± 4.7 t C-CO_2 ha^{-1} a^{-1}). The model cannot determine whether the system is net gaining or losing carbon. The annual mean and temporal dynamics of the modelled emissions are well in accordance with the measured emissions. The effect of grazing has a minor influence on the total ecosystem respiration (Figure III.7), resulting in a wider range of both measured and modelled emissions. Grazing, i.e., the reduction of root biomass, results in two contrary processes: a reduction in maintenance respiration and an increase in autotrophic respiration (Raich and Tufekciogul, 2000).

Forest C cycle: The forest model predicts an annual C input of 8.9 ± 0.6 t C-CO_2 ha^{-1} a^{-1}, which is quite low compared to the estimations for old-growth beech forests in Europe, with reported rates

from 14.4 to 18.3 t C-CO_2 ha^{-1} a^{-1} (Molina-Herrera et al., 2015). However, C uptake rates vary in magnitude, with values ranging from 3 to 34 t C-CO_2 ha^{-1} a^{-1} for different forests in different growing stages (Waring et al., 1998). As our study site is a mixture of young and old beech trees, we assume that it has 40 - 50% less biomass compared to an old beech forest. Of the modelled C input, 6.6 ± 0.5 t C-CO_2 ha^{-1} a^{-1} leaves the system as gaseous CO_2. The rest is accumulated in the biomass and soil. The annual mean and dynamics of the modelled emissions are in accordance with the measured emissions. We expected to see rising emissions with litter fall in autumn (Raich and Tufekciogul, 2000), but cannot report this effect, either with measurements or with model results (Figure III.6).

Conclusion

We presented a two-year measurement campaign of trace gas emissions from adjacent land uses i.e., arable land, grassland and forest ecosystems, with concurrent model development and rigorous testing through a model-data fusion.

We found high emissions of N_2O and CO_2 on our arable land sites, low emissions on grassland sites and the lowest emissions on the forest sites. These observations enable us to investigate the underlying effects of plant growth, temperature and WFPS, land use effects, seasonal patterns and management effects. Respiration amounts rise in less shaded (warmer) areas of the forest, while N_2O emissions increase towards the foothills of the forest and arable land sites due to nitrogen accumulation. Highly variable N_2O emissions in forests resulted in large uncertainties in the model verification data, which translated into large uncertainties in the model results for forests.

Table III.6. Overall posterior model performance of LandscapeDNDC on different land uses in reproducing GHG emission data. Subjectively classified into (1) good, (2) medium and (3) poor model performance in simulating reliable annual sums, seasonal patterns and magnitudes of management events (e.g., fertilizer application). NA = not applicable, i.e., no forest management during modelled period from 2010-2016.

Modelled performance on each land use	N_2O emissions			CO_2 emissions		
	annual	seasonal	management	annual	seasonal	management
Arable land (A1-3)	2	1	1	1	2	3
Grassland (G1)	1	2	1	1	1	1
Forest (W1-W3)	2	2	NA	1	2	NA

Detailed measured data on soil and management allowed us to fit the biogeochemical model LandscapeDNDC to the measured soil moisture, yield and GHG emissions of CO_2 and N_2O. A subjective conclusion about the overall model performance is shown in Table III.6: The model reproduced the measured data reasonably well in time, separated into seasons and management events. The model performance was best in predicting management effects on N_2O emissions and annual CO_2 emissions for all land uses. With regard to land use, the simulations for grassland sites

work best, followed by those for arable land. The simulations for N_2O emissions on arable land outperform those for CO_2, and vice versa for grassland. Low emissions on forest sites were generally difficult to depict using our modelling approach.

The model-data fusion approach allowed us to identify model structural deficiencies that would likely increase model performances if addressed in Landscape DNDC: missing N_2O uptake processes; missing NO_3^- (and potentially dissolved organic nitrogen) uptake through shallow groundwater; missing lateral interaction on hillslopes due to the 1D model setup.

Furthermore, posterior model runs allowed for the quantification of the magnitude and uncertainty of unmeasured C and N cycle fluxes. The investigated forest site generally acts as the largest sink for C and N, with annual sequestration rates of 2.4 ± 1.1 t C ha^{-1} and 3.3 ± 2.0 kg N ha^{-1}. Whether the extensive grazed grassland is also acting as a sink for C with 1.4 ± 4.7 t C ha^{-1} per year remains uncertain, while the N cycle of the grassland model cannot be closed with the given settings. Shrinking N soil pools indicate a missing input, which we assume to be shallow groundwater with an additional N supply of approximately 40.9 ± 25.9 kg N ha^{-1} a^{-1}.

Current land use in this catchment is dominated by forests (37%) and arable land (35%), whereas grassland sites (11%) are mainly distributed along the stream. From the viewpoint of climate-smart landscapes, the measured data suggest the benefit of forests in a landscape, as they have the fewest GHG emissions. Riparian zones can act as sinks of N but only during the vegetation period and during times when roots have access to groundwater. Arable land use produces high amounts of N_2O, not throughout the year, but rather, in spring after fertilizer application.

Potential interactions of land use patterns cannot be quantified with the current one-dimensional model approach. However, the dataset could be used in future studies to quantify the nitrate uptake of riparian zones in more detail, e.g., by coupling LandscapeDNDC to a hydrological model, as done by Klatt et al. (2017). Such a model setup would also allow for upscaling in space, e.g., for the generation of GHG inventories or an analysis of more detailed management scenarios in time.

Acknowledgements. We acknowledge the financial support provided by the Deutsche Forschungsgemeinschaft (DFG) for Tobias Houska (BR2238/13-1). Special thanks deserves Felix Kruck, Eva Holthof and Michael Herzog for their fieldwork during any weather conditions, Anja Schaefler-Schmid and Julia Valverde for lab analysis and providing the chamber sampling equipment as well as the farmer, for letting us study his land and providing the detailed management information.

References

Ackley, D. H.: An empirical study of bit vector function optimization, Genet. Algorithms Simulated Annealing, 1, 170–204, 1987.

Alfi, A.: PSO with Adaptive Mutation and Inertia Weight and Its Application in Parameter Estimation of Dynamic Systems, Acta Autom. Sin., 37(5), 541–549, doi:10.1016/S1874-1029(11)60205-X, 2011.

Anon: The Open Source Definition | Open Source Initiative, [online] Available from: http://opensource.org/docs/osd (Accessed 20 July 2015).

Arias-Navarro, C., Díaz-Pinés, E., Kiese, R., Rosenstock, T. S., Rufino, M. C., Stern, D., Neufeldt, H., Verchot, L. V. and Butterbach-Bahl, K.: Gas pooling: A sampling technique to overcome spatial heterogeneity of soil carbon dioxide and nitrous oxide fluxes, Soil Biol. Biochem., 67, 20–23, doi:10.1016/j.soilbio.2013.08.011, 2013.

Bárdossy, A. and Singh, S. K.: Robust estimation of hydrological model parameters, Hydrol. Earth Syst. Sci. Discuss., 5(3), 1641–1675, doi:10.5194/hess-12-1273-2008, 2008.

Barton, L., Kiese, R., Gatter, D., Butterbach-Bahl, K., Buck, R., Hinz, C. and Murphy, D. V.: Nitrous oxide emissions from a cropped soil in a semi-arid climate, Glob. Change Biol., 14(1), 177–192, doi:10.1111/j.1365-2486.2007.01474.x, 2008.

Behrangi, A., Khakbaz, B., Vrugt, J. A., Duan, Q. and Sorooshian, S.: Comment on "Dynamically dimensioned search algorithm for computationally efficient watershed model calibration" by Bryan A. Tolson and Christine A. Shoemaker, Water Resour. Res., 44(12), W12603, doi:10.1029/2007WR006429, 2008.

Bergström, S., Singh, V. P. and others: The HBV model., Comput. Models Watershed Hydrol., 443–476, 1995.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320(1–2), 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K.: On the concept of model structural error, Proc. Int. Workshop Uncertain. Precaut. Environ. Model., 52(6), 167–175, 2007.

Beven, K.: Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrol. Sci. J., doi:10.1080/02626667.2015.1031761, 2015.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6(3), 279–298, doi:10.1002/hyp.3360060305, 1992.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249(1–4), 11–29, doi:10.1016/S0022-1694(01)00421-8, 2001.

Blagodatsky, S. A. and Richter, O.: Microbial growth in soil and nitrogen turnover: a theoretical model considering the activity state of microorganisms, Soil Biol. Biochem., 30(13), 1743–1755, doi:10.1016/S0038-0717(98)00028-5, 1998.

Bloom, A. A. and Williams, M.: Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological "common sense" in a model-data fusion framework, Biogeosciences, 12(5), 1299, doi:10.5194/bg-12-1299-2015, 2015.

Boden, T. A., Marland, G. and Andres, R. J.: Global, regional and national fossil fuel co2 emissions. [online] Available from: http://cdiac.ornl.gov/trends/emis/overview_2007.html (Accessed 17 August 2016), 2010.

Bormann, H., Breuer, L., Gräff, T. and Huisman, J. A.: Analysing the effects of soil properties changes associated with land use changes on the simulated water balance: A comparison of three hydrological catchment models for scenario analysis, Ecol. Model., 209(1), 29–40, doi:10.1016/j.ecolmodel.2007.07.004, 2007.

Bouwman, A. F., Boumans, L. J. M. and Batjes, N. H.: Emissions of N2O and NO from fertilized fields: Summary of available measurement data, Glob. Biogeochem. Cycles, 16(4), 1–13, doi:10.1029/2001GB001811, 2002.

Bouwman, A. F., Stehfest, E. and van Kessel, C.: Nitrous oxide emissions from the nitrogen cycle in arable agriculture: Estimation and mitigation, Nitrous Oxide Clim. Change, 85–106, 2010.

ter Braak, C. J. and Vrugt, J. A.: Differential evolution Markov chain with snooker updater and fewer chains, Stat. Comput., 18(4), 435–446, doi:10.1007/s11222-008-9104-9, 2008.

Bremner, J. M.: Sources of nitrous oxide in soils, Nutr. Cycl. Agroecosystems, 49(1–3), 7–16, 1997.

Breuer, L., Kiese, R. and Butterbach-Bahl, K.: Temperature and moisture effects on nitrification rates in tropical rain-forest soils, Soil Sci. Soc. Am. J., 66(3), 834–844, doi:10.2136/sssaj2002.8340, 2002.

Bruijn, A. M. G. de and Butterbach-Bahl, K.: Linking carbon and nitrogen mineralization with microbial responses to substrate availability — the DECONIT model, Plant Soil, 328(1–2), 271–290, doi:10.1007/s11104-009-0108-9, 2009.

Bruijn, A. M. G. de, Grote, R. and Butterbach-Bahl, K.: An alternative modelling approach to predict emissions of N2O and NO from forest soils, Eur. J. For. Res., 130(5), 755–773, doi:10.1007/s10342-010-0468-y, 2011.

Burton, D. L., Li, X. and Grant, C. A.: Influence of fertilizer nitrogen source and management practice on N2O emissions from two Black Chernozemic soils, Can. J. Soil Sci., 88(2), 219–227, doi:10.4141/CJSS06020, 2008.

Butterbach-Bahl, K. and Dannenmann, M.: Denitrification and associated soil N2O emissions due to agricultural activities in a changing climate, Curr. Opin. Environ. Sustain., 3(5), 389–395, doi:10.1016/j.cosust.2011.08.004, 2011.

Butterbach-Bahl, K., Kahl, M., Mykhayliv, L., Werner, C., Kiese, R. and Li, C.: A European-wide inventory of soil NO emissions using the biogeochemical models DNDC/Forest-DNDC, Atmos. Environ., 43(7), 1392–1402, doi:10.1016/j.atmosenv.2008.02.008, 2009.

Butterbach-Bahl, K., Baggs, E. M., Dannenmann, M., Kiese, R. and Zechmeister-Boltenstern, S.: Nitrous oxide emissions from soils: how well do we understand the processes and their controls?,

Philos. Trans. R. Soc. Lond. B Biol. Sci., 368(1621), 20130122, doi:10.1098/rstb.2013.0122, 2013.

Buyanovsky, G. A., Wagner, G. H. and Gantzer, C. J.: Soil respiration in a winter wheat ecosystem, Soil Sci Soc Am J, 50(2), 338–344, 1986.

Buytaert, W., Reusser, D., Krause, S. and Renaud, J.-P.: Why can't we do better than Topmodel?, Hydrol. Process., 22(20), 4175–4179, doi:10.1002/hyp.7125, 2008.

Chang, J. F., Viovy, N., Vuichard, N., Ciais, P., Wang, T., Cozic, A., Lardy, R., Graux, A.-I., Klumpp, K., Martin, R. and others: Incorporating grassland management in ORCHIDEE: model description and evaluation at 11 eddy-covariance sites in Europe, Geosci. Model Dev., 6(6), 2165–2181, doi:10.5194/gmd-6-2165-2013, 2013.

Clark, M. P., Kavetski, D. and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resour. Res., 47(9), W09301, doi:10.1029/2010WR009827, 2011.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J. and Rasmussen, R. M.: A unified approach for process-based hydrologic modeling: 1. Modeling concept, Water Resour. Res., 51(4), 2498–2514, doi:10.1002/2015WR017198, 2015.

Cole, C. V., Duxbury, J., Freney, J., Heinemeyer, O., Minami, K., Mosier, A., Paustian, K., Rosenberg, N., Sampson, N. and Sauerbeck, D.: Global estimates of potential mitigation of greenhouse gas emissions by agriculture, Nutr. Cycl. Agroecosystems, 49(1–3), 221–228, doi:10.1023/A:1009731711346, 1997.

Cui, F., Zheng, X., Liu, C., Wang, K., Zhou, Z. and Deng, J.: Assessing biogeochemical effects and best management practice for a wheat–maize cropping system using the DNDC model, Biogeosciences, 11(1), 91–107, doi:10.5194/bg-11-91-2014, 2014.

Dalcín, L., Paz, R., Storti, M. and D'Elía, J.: MPI for Python: Performance improvements and MPI-2 extensions, J. Parallel Distrib. Comput., 68(5), 655–662, doi:10.1016/j.jpdc.2007.09.005, 2008.

De Bruijn, A. M. G., Butterbach-Bahl, K., Blagodatsky, S. and Grote, R.: Model evaluation of different mechanisms driving freeze–thaw N2O emissions, Agric. Ecosyst. Environ., 133(3), 196–207, doi:10.1016/j.agee.2009.04.023, 2009.

Dirnböck, T., Kobler, J., Kraus, D., Grote, R. and Kiese, R.: Impacts of management and climate change on nitrate leaching in a forested karst area, J. Environ. Manage., 165, 243–252, doi:10.1016/j.jenvman.2015.09.039, 2016.

Doherty, J. and Johnston, J. M.: Methodologies for Calibration and Predictive Analysis of a Watershed Model1, JAWRA J. Am. Water Resour. Assoc., 39(2), 251–265, doi:10.1111/j.1752-1688.2003.tb04381.x, 2003.

Duan, Q., Sorooshian, S. and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28(4), 1015–1031, doi:10.1029/91WR02985, 1992.

Duan, Q., Sorooshian, S. and Gupta, V. K.: Optimal use of the SCE-UA global optimization method for calibrating watershed models, J. Hydrol., 158(3), 265–284, 1994.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrol. Sci. J., 55(1), 58–78, doi:10.1080/02626660903526292, 2010.

Firestone, M. K. and Davidson, E. A.: Microbiological basis of NO and N2O production and consumption in soil, Exch. Trace Gases Terr. Ecosyst. Atmosphere, 47, 7–21, 1989.

Flechard, C. R., Neftel, A., Jocher, M., Ammann, C. and Fuhrer, J.: Bi-directional soil/atmosphere N2O exchange over two mown grassland systems with contrasting management practices, Glob. Change Biol., 11(12), 2114–2127, doi:10.1111/j.1365-2486.2005.01056.x, 2005.

Flechard, C. R., Nemitz, E., Smith, R. I., Fowler, D., Vermeulen, A. T., Bleeker, A., Erisman, J. W., Simpson, D., Zhang, L. and Tang, Y. S.: Dry deposition of reactive nitrogen to European ecosystems: a comparison of inferential models across the NitroEurope network, Atmospheric Chem. Phys., 11(6), 2703–2728, doi:10.5194/acp-11-2703-2011, 2011.

Frolking, S. E., Mosier, A. R., Ojima, D. S., Li, C., Parton, W. J., Potter, C. S., Priesack, E., Stenger, R., Haberbosch, C., Dörsch, P., Flessa, H. and Smith, K. A.: Comparison of N2O emissions from soils at three temperate agricultural sites: simulations of year-round measurements by four models, Nutr. Cycl. Agroecosystems, 52(2–3), 77–105, doi:10.1023/A:1009780109748, 1998.

Gabrielle, B., Laville, P., Duval, O., Nicoullaud, B., Germon, J.-C. and Hénault, C.: Process-based modeling of nitrous oxide emissions from wheat-cropped soils at the subregional scale, Glob. Biogeochem. Cycles, 20(4), GB4018, doi:10.1029/2006GB002686, 2006.

Gao, M., Qiu, J., Li, C., Wang, L., Li, H. and Gao, C.: Modeling nitrogen loading from a watershed consisting of cropland and livestock farms in China using Manure-DNDC, Agric. Ecosyst. Environ., 185, 88–98, doi:10.1016/j.agee.2013.10.023, 2014.

Gelman, A. and Rubin, D. B.: Inference from iterative simulation using multiple sequences, Stat. Sci., 7(4), 457–472, 1992.

Geweke, J.: Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments, in IN BAYESIAN STATISTICS, pp. 169–193, University Press., 1992.

Gilhespy, S. L., Anthony, S., Cardenas, L., Chadwick, D., del Prado, A., Li, C., Misselbrook, T., Rees, R. M., Salas, W., Sanz-Cobena, A., Smith, P., Tilston, E. L., Topp, C. F. E., Vetter, S. and Yeluripati, J. B.: First 20 years of DNDC (DeNitrification DeComposition): Model evolution, Ecol. Model., 292, 51–62, doi:10.1016/j.ecolmodel.2014.09.004, 2014.

Gilmanov, T. G., Soussana, J. F., Aires, L., Allard, V., Ammann, C., Balzarolo, M., Barcza, Z., Bernhofer, C., Campbell, C. L. and Cernusca, A.: Partitioning European grassland net ecosystem CO 2 exchange into gross primary productivity and ecosystem respiration using light response function analysis, Agric. Ecosyst. Environ., 121(1), 93–120, doi:10.1016/j.agee.2006.12.008, 2007.

Giltrap, D. L., Li, C. and Saggar, S.: DNDC: A process-based model of greenhouse gas fluxes from agricultural soils, Agric. Ecosyst. Environ., 136(3–4), 292–300, doi:10.1016/j.agee.2009.06.014, 2010.

Glatzel, S. and Stahr, K.: Methane and nitrous oxide exchange in differently fertilised grassland in southern Germany, Plant Soil, 231(1), 21–35, doi:10.1023/A:1010315416866, 2001.

Gong, Y. M., Mohammat, A., Liu, X. J., Li, K. H., Christie, P., Fang, F., Song, W., Chang, Y. H., Han, W. X., Lü, X. T., Liu, Y. Y. and Hu, Y. K.: Response of carbon dioxide emissions to sheep grazing and N application in an alpine grassland – Part 1: Effect of sheep grazing, Biogeosciences, 11(7), 1743–1750, doi:10.5194/bg-11-1743-2014, 2014.

Gong, Z.: Estimation of mixed Weibull distribution parameters using the SCEM-UA algorithm: Application and comparison with MLE in automotive reliability analysis, Reliab. Eng. Syst. Saf., 91(8), 915–922, doi:10.1016/j.ress.2005.09.007, 2006.

Goodman, J. and Weare, J.: Ensemble samplers with affine invariance, Commun. Appl. Math. Comput. Sci., 5(1), 65–80, doi:10.2140/camcos.2010.5.65, 2010.

Goswami, S., Gamon, J., Vargas, S. and Tweedie, C.: Relationships of NDVI, Biomass, and Leaf Area Index (LAI) for six key plant species in Barrow, Alaska, PeerJ PrePrints., 2015.

Graedel, T. E. and Crutzen, P. J.: The changing atmosphere., Sci. Am., 58–68, 1989.

Griewank, A. O.: Generalized descent for global optimization, J. Optim. Theory Appl., 34(1), 11–39, 1981.

Grote, R., Lehmann, E., Brümmer, C., Brüggemann, N., Szarzynski, J. and Kunstmann, H.: Modelling and observation of biosphere–atmosphere interactions in natural savannah in Burkina Faso, West Africa, Phys. Chem. Earth Parts ABC, 34(4–5), 251–260, doi:10.1016/j.pce.2008.05.003, 2009.

Guillaume, J. H., Kummu, M., Räsänen, T. A. and Jakeman, A. J.: Prediction under uncertainty as a boundary problem: A general formulation using Iterative Closed Question Modelling, Environ. Model. Softw., 70, 97–112, doi:10.1016/j.envsoft.2015.04.004, 2015.

Guinot, V., Cappelaere, B., Delenne, C. and Ruelland, D.: Towards improved criteria for hydrological model calibration: theoretical analysis of distance- and weak form-based functions, J. Hydrol., 401(1–2), 1–13, doi:10.1016/j.jhydrol.2011.02.004, 2011.

Haan, C. T., Storm, D. E., Al-Issa, T., Prabhu, S., Sabbagh, G. J. and Edwards, D. R.: Effect of parameter distributions on uncertainty analysis of hydrologic models, Trans. Asae, 41(1), 65–70, 1998.

Haas, E., Klatt, S., Fröhlich, A., Kraft, P., Werner, C., Kiese, R., Grote, R., Breuer, L. and Butterbach-Bahl, K.: LandscapeDNDC: a process model for simulation of biosphere–atmosphere–hydrosphere exchange processes at site and regional scale, Landsc. Ecol., 28(4), 615–636, 2013.

Harp, D. R. and Vesselinov, V. V.: An agent-based approach to global uncertainty and sensitivity analysis, Comput. Geosci., 40, 19–27, doi:10.1016/j.cageo.2011.06.025, 2012.

He, J., Jones, J. W., Graham, W. D. and Dukes, M. D.: Influence of likelihood function choice for estimating crop model parameters using the generalized likelihood uncertainty estimation method, Agric. Syst., 103(5), 256–264, doi:10.1016/j.agsy.2010.01.006, 2010.

Hoffman, M. D. and Gelman, A.: The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, ArXiv11114246 Cs Stat [online] Available from: http://arxiv.org/abs/1111.4246 (Accessed 27 January 2015), 2011.

Holst, J., Liu, C., Yao, Z., Brüggemann, N., Zheng, X., Giese, M. and Butterbach-Bahl, K.: Fluxes of nitrous oxide, methane and carbon dioxide during freezing–thawing cycles in an Inner Mongolian steppe, Plant Soil, 308(1–2), 105–117, doi:10.1007/s11104-008-9610-8, 2008.

Hossain, F., Anagnostou, E. N. and Bagtzoglou, A. C.: On Latin Hypercube sampling for efficient uncertainty estimation of satellite rainfall observations in flood prediction, Comput. Geosci., 32(6), 776–792, doi:10.1016/j.cageo.2005.10.006, 2006.

Houska, T., Multsch, S., Kraft, P., Frede, H.-G. and Breuer, L.: Monte Carlo-based calibration and uncertainty analysis of a coupled plant growth and hydrological model, Biogeosciences, 11(7), 2069–2082, doi:10.5194/bg-11-2069-2014, 2014.

Houska, T., Kraft, P., Chamorro-Chavez, A. and Breuer, L.: SPOTting Model Parameters Using a Ready-Made Python Package, PLoS ONE, 10(12), e0145180, doi:10.1371/journal.pone.0145180, 2015.

Houska, T., Kraus, D., Kiese, R. and Breuer, L.: Constraining a complex biogeochemical model for multi-site greenhouse gas emission simulations by model-data fusion, Biogeosciences Discuss, 2017, 1–28, doi:10.5194/bg-2017-96, 2017a.

Houska, T., Kraft, P., Liebermann, R., Klatt, S., Kraus, D., Haas, E., Santabárbara, I., Kiese, R., Butterbach-Bahl, K., Müller, C. and Breuer, L.: Rejecting hydro-biogeochemical model structures by multi-criteria evaluation, Environ. Model. Softw., 93, 1–12, doi:10.1016/j.envsoft.2017.03.005, 2017b.

Huang, M. and Liang, X.: On the assessment of the impact of reducing parameters and identification of parameter uncertainties for a hydrologic model with applications to ungauged basins, J. Hydrol., 320(1–2), 37–61, doi:10.1016/j.jhydrol.2005.07.010, 2006.

IPCC: Revised 1996 IPCC guidelines for national greenhouse gas inventories. v. 1: Greenhouse gas inventory reporting instructions.-v. 2: Greenhouse gas inventory workbook.-v. 3: Greenhouse gas inventory reference manual, 1997.

IPCC: Contribution of working group III to the fourth assessment report of the intergovernmental panel on climate change, 2007.

Jäger, H.-J., Schmidt, S. W., Kammann, C., Grünhage, L., Müller, C. and Hanewald, K.: The University of Giessen Free-Air Carbon dioxide Enrichment study: Description of the experimental site and of a new enrichment system, J. Appl. Bot., 77(5–6), 117–127, 2003.

Janssens, I. A., Lankreijer, H., Matteucci, G., Kowalski, A. S., Buchmann, N., Epron, D., Pilegaard, K., Kutsch, W., Longdoz, B. and Grünwald, T.: Productivity overshadows temperature in

determining soil and ecosystem respiration across European forests, Glob. Change Biol., 7(3), 269–278, doi:10.1046/j.1365-2486.2001.00412.x, 2001.

Jansson, P.-E.: CoupModel: model use, calibration, and validation, Trans. ASABE, 55(4), 1337–1344, doi:10.13031/2013.42245, 2012.

Jung, B. S., Karnev, B. W. and Lambert, M. F.: Benchmark tests of evolutionary algorithms: mathematic evaluation and application to water distribution systems, J. Environ. Inform., 7(1), 24–35, doi:10.3808/jei.200600064, 2006.

Jungkunst, H. F., Freibauer, A., Neufeldt, H. and Bareth, G.: Nitrous oxide emissions from agricultural land use in Germany—a synthesis of available annual field data, J. Plant Nutr. Soil Sci., 169(3), 341–351, doi:10.1002/jpln.200521954, 2006.

Kammann, C., Grünhage, L., Müller, C., Jacobi, S. and Jäger, H.-J.: Seasonal variability and mitigation options for N2O emissions from differently managed grasslands, Environ. Pollut., 102, 179–186, doi:10.1016/S0269-7491(98)80031-6, 1998.

Kammann, C., Grünhage, L. and Jäger, H.-J.: A new sampling technique to monitor concentrations of CH4, N2O and CO2 in air at well-defined depths in soils with varied water potential, Eur. J. Soil Sci., 52(2), 297–303, doi:10.1046/j.1365-2389.2001.00380.x, 2001.

Kammann, C., Müller, C., Grünhage, L. and Jäger, H.-J.: Elevated CO2 stimulates N2O emissions in permanent grassland, Soil Biol. Biochem., 40(9), 2194–2205, doi:10.1016/j.soilbio.2008.04.012, 2008.

Karaboga, D. and Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm, J. Glob. Optim., 39(3), 459–471, doi:10.1007/s10898-007-9149-x, 2007.

Kavetski, D., Kuczera, G. and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resour. Res., 42(3), W03407, doi:10.1029/2005WR004368, 2006.

Keenan, T. F., Carbone, M. S., Reichstein, M. and Richardson, A. D.: The model–data fusion pitfall: assuming certainty in an uncertain world, Oecologia, 167(3), 587, doi:10.1007/s00442-011-2106-x, 2011.

Kiese, R. and Butterbach-Bahl, K.: N2O and CO2 emissions from three different tropical forest sites in the wet tropics of Queensland, Australia, Soil Biol. Biochem., 34(7), 975–987, doi:10.1016/S0038-0717(02)00031-7, 2002.

Kiese, R., Heinzeller, C., Werner, C., Wochele, S., Grote, R. and Butterbach-Bahl, K.: Quantification of nitrate leaching from German forest ecosystems by use of a process oriented biogeochemical model, Environ. Pollut., 159(11), 3204–3214, doi:10.1016/j.envpol.2011.05.004, 2011.

Kim, Y., Seo, Y., Kraus, D., Klatt, S., Haas, E., Tenhunen, J. and Kiese, R.: Estimation and mitigation of N2O emission and nitrate leaching from intensive crop cultivation in the Haean catchment, South Korea, Sci. Total Environ., 529, 40–53, doi:10.1016/j.scitotenv.2015.04.098, 2015.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. and others: Optimization by simmulated annealing, science, 220(4598), 671–680, 1983.

Kitanidis, P. K. and Lane, R. W.: Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss-Newton method, J. Hydrol., 79(1–2), 53–71, doi:10.1016/0022-1694(85)90181-7, 1985.

Klatt, S., Kraus, D., Kraft, P., Breuer, L., Wlotzka, M., Heuveline, V., Haas, E., Kiese, R. and Butterbach-Bahl, K.: Exploring impacts of vegetated buffer strips on nitrogen cycling using a spatially explicit hydro-biogeochemical modeling approach, Environ. Model. Softw., 90, 55–67, doi:10.1016/j.envsoft.2016.12.002, 2017.

Kraft, P., Vaché, K. B., Frede, H.-G. and Breuer, L.: CMF: A Hydrological Programming Language Extension For Integrated Catchment Models, Environ. Model. Softw., 26(6), 828–830, doi:10.1016/j.envsoft.2010.12.009, 2011.

Kraft, P., Haas, E., Klatt, S., Kiese, R., Butterbach-Bahl, K., Frede, H.-G. and Breuer, L.: Modelling nitrogen transport and turnover at the hillslope scale–a process oriented approach, in AGU Fall Meeting Abstracts, vol. 1, p. 0688. [online] Available from: http://www.iemss.org/iemss2012/proceedings/F3_0872_Kraft_et_al.pdf (Accessed 8 January 2015), 2012.

Kraus, D., Weller, S., Klatt, S., Haas, E., Wassmann, R., Kiese, R. and Butterbach-Bahl, K.: A new LandscapeDNDC biogeochemical module to predict CH4 and N2O emissions from lowland rice and upland cropping systems, Plant Soil, 386(1–2), 125–149, doi:10.1007/s11104-014-2255-x, 2015.

Kraus, D., Weller, S., Klatt, S., Santabárbara, I., Haas, E., Wassmann, R., Werner, C., Kiese, R. and Butterbach-Bahl, K.: How well can we assess impacts of agricultural land management changes on the total greenhouse gas balance (CO 2, CH 4 and N 2 O) of tropical rice-cropping systems with a biogeochemical model?, Agric. Ecosyst. Environ., 224, 104–115, doi:10.1016/j.agee.2016.03.037, 2016.

Krauße, T. and Cullmann, J.: Towards a more representative parametrisation of hydrologic models via synthesizing the strengths of Particle Swarm Optimisation and Robust Parameter Estimation, Hydrol. Earth Syst. Sci., 16(2), 603–629, doi:10.5194/hess-16-603-2012, 2012.

Kröbel, R., Sun, Q., Ingwersen, J., Chen, X., Zhang, F., Müller, T. and Römheld, V.: Modelling water dynamics with DNDC and DAISY in a soil of the North China Plain: a comparative study, Environ. Model. Softw., 25(4), 583–601, doi:10.1016/j.envsoft.2009.09.003, 2010.

LandscapeDNDC: LandscapeDNDC: a regional-scale DNDC-based process model, [online] Available from: http://svn.gap.fzk.de/ (Accessed 29 September 2015), 2015.

Lazzarotto, P., Calanca, P. and Fuhrer, J.: Dynamics of grass–clover mixtures—an analysis of the response to management with the PROductive GRASsland Simulator (PROGRASS), Ecol. Model., 220(5), 703–724, doi:10.1016/j.ecolmodel.2008.11.023, 2009.

Leffelaar, P. A. and Wessel, W. W.: Denitrification in a homogeneous, closed system: experiment and simulation., Soil Sci., 146(5), 335–349, 1988.

Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35(1), 233–241, doi:10.1029/1998WR900018, 1999.

Lehuger, S., Gabrielle, B., Oijen, M. van, Makowski, D., Germon, J.-C., Morvan, T. and Hénault, C.: Bayesian calibration of the nitrous oxide emission module of an agro-ecosystem model, Agric. Ecosyst. Environ., 133(3–4), 208–222, doi:10.1016/j.agee.2009.04.022, 2009.

Li, C., Frolking, S. and Frolking, T. A.: A model of nitrous oxide evolution from soil driven by rainfall events: 1. Model structure and sensitivity, J Geophys Res, 97(D9), 9759–9776, doi:10.1029/92JD00509, 1992.

Li, C., Aber, J., Stange, F., Butterbach-Bahl, K. and Papen, H.: A process-oriented model of N2O and NO emissions from forest soils: 1. Model development, J.Geophys.Res., 105, 4369–4384, doi:10.1029/1999JD900949, 2000.

Li, L., Xia, J., Xu, C.-Y. and Singh, V. P.: Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models, J. Hydrol., 390(3–4), 210–221, doi:10.1016/j.jhydrol.2010.06.044, 2010.

Liebermann, R., Houska, T., Kraft, P., Klatt, S., Kraus, D., Haas, E., Müller, C. and Breuer, L.: Closing the N-budget: How Simulated Groundwater-borne Nitrate Supply Affects Plant Growth and Greenhouse Gas Emissions on a Temperate Grassland, PLoS ONE, submitted, 2017.

Liu, L. and Greaver, T. L.: A review of nitrogen enrichment effects on three biogenic GHGs: the CO2 sink may be largely offset by stimulated N2O and CH4 emission, Ecol. Lett., 12(10), 1103–1117, doi:10.1111/j.1461-0248.2009.01351.x, 2009.

Lohila, A., Aurela, M., Regina, K. and Laurila, T.: Soil and total ecosystem respiration in agricultural fields: effect of soil and crop type, Plant Soil, 251(2), 303–317, doi:10.1023/A:1023004205844, 2003.

Ludwig, B., Jäger, N., Priesack, E. and Flessa, H.: Application of the DNDC model to predict N2O emissions from sandy arable soils with differing fertilization in a long-term experiment, J. Plant Nutr. Soil Sci., 174(3), 350–358, doi:10.1002/jpln.201000040, 2011.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D.: WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility, Stat. Comput., 10(4), 325–337, doi:10.1023/A:1008929526011, 2000.

Madsen, H., Wilson, G. and Ammentorp, H. C.: Comparison of different automated strategies for calibration of rainfall-runoff models, J. Hydrol., 261(1–4), 48–59, doi:10.1016/S0022-1694(01)00619-9, 2002.

Matott, L. S., Babendreier, J. E. and Purucker, S. T.: Evaluating uncertainty in integrated environmental models: A review of concepts and tools, Water Resour. Res., 45(6), W06421, doi:10.1029/2008WR007301, 2009.

Matott, L. S., Hymiak, B., Reslink, C., Baxter, C. and Aziz, S.: Telescoping strategies for improved parameter estimation of environmental simulation models, Comput. Geosci., 60, 156–167, doi:10.1016/j.cageo.2013.07.023, 2013.

McKay, M. D., Beckman, R. J. and Conover, W. J.: Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, Technometrics, 21(2), 239–245, doi:10.1080/00401706.1979.10489755, 1979.

McMillan, H., Krueger, T. and Freer, J.: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality, Hydrol. Process., 26(26), 4078–4111, doi:10.1002/hyp.9384, 2012.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E.: Equation of state calculations by fast computing machines, J. Chem. Phys., 21(6), 1087–1092, 1953.

Metzger, C. M. H., Heinichen, J., Eickenscheidt, T. and Drösler, M.: Impact of land-use intensity on the relationships between vegetation indices, photosynthesis and biomass of intensively and extensively managed grassland fens, Grass Forage Sci., 72(1), 50–63, doi:10.1111/gfs.12223, 2017.

Moeck, C., Hunkeler, D. and Brunner, P.: Tutorials as a flexible alternative to GUIs: An example for advanced model calibration using Pilot Points, Environ. Model. Softw., 66, 78–86, doi:10.1016/j.envsoft.2014.12.018, 2015.

Molina-Herrera, S., Grote, R., Santabárbara-Ruiz, I., Kraus, D., Klatt, S., Haas, E., Kiese, R. and Butterbach-Bahl, K.: Simulation of CO2 Fluxes in European Forest Ecosystems with the Coupled Soil-Vegetation Process Model "LandscapeDNDC," Forests, 6(6), 1779–1809, doi:10.3390/f6061779, 2015.

Molina-Herrera, S., Haas, E., Klatt, S., Kraus, D., Augustin, J., Magliulo, V., Tallec, T., Ceschia, E., Ammann, C. and Loubet, B.: A modeling study on mitigation of N2O emissions and NO3 leaching at different agricultural sites across Europe using LandscapeDNDC, Sci. Total Environ., 553, 128–140, doi:10.1016/j.scitotenv.2015.12.099, 2016.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D. and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Trans Asabe, 50(3), 885–900, doi:10.13031/2013.23153, 2007.

Mualem, Y.: A New Model for Predicting the Hydraulic Conductivity of Unsaturated Porous Media, Water Resour. Res., 12, 513–522, doi:10.1029/WR012i003p00513, 1976.

Müller, C., Martin, M., Stevens, R. J., Laughlin, R. J., Kammann, C., Ottow, J. C. G. and Jäger, H.-J.: Processes leading to N 2 O emissions in grassland soil during freezing and thawing, Soil Biol. Biochem., 34(9), 1325–1331, doi:10.1016/S0038-0717(02)00076-7, 2002.

Murphy, C., Fealy, R., Charlton, R. and Sweeney, J.: The reliability of an 'off-the-shelf' conceptual rainfall runoff model for use in climate impact assessment: uncertainty quantification using Latin hypercube sampling, Area, 38(1), 65–78, doi:10.1111/j.1475-4762.2006.00656.x, 2006.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D. and Mendoza, B.: Anthropogenic and natural radiative forcing, IPCC WGI Fifth Assess. Rep., 423, 2013.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, J. Hydrol., 10(3), 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Necpálová, M., Anex, R. P., Fienen, M. N., Del Grosso, S. J., Castellano, M. J., Sawyer, J. E., Iqbal, J., Pantoja, J. L. and Barker, D. W.: Understanding the DayCent model: Calibration, sensitivity,

and identifiability through inverse modeling, Environ. Model. Softw., 66, 110–130, doi:10.1016/j.envsoft.2014.12.011, 2015.

Neftel, A., Flechard, C., Ammann, C., Conen, F., Emmenegger, L. and Zeyer, K.: Experimental assessment of N2O background fluxes in grassland systems, Tellus B, 59(3), 470–482, doi:10.1111/j.1600-0889, 2007.

Nylinder, J., Stenberg, M., Jansson, P.-E., Klemedtsson, Å. K., Weslien, P. and Klemedtsson, L.: Modelling uncertainty for nitrate leaching and nitrous oxide emissions based on a Swedish field experiment with organic crop rotation, Agric. Ecosyst. Environ., 141(1–2), 167–183, doi:10.1016/j.agee.2011.02.027, 2011.

Oenema, O., Velthof, G. L., Yamulki, S. and Jarvis, S. C.: Nitrous oxide emissions from grazed grassland, Soil Use Manag., 13(s4), 288–295, doi:10.1111/j.1475-2743, 1997.

Oijen, M. V., Rougier, J. and Smith, R.: Bayesian calibration of process-based forest models: bridging the gap between models and data, Tree Physiol., 25(7), 915–927, doi:10.1093/treephys/25.7.915, 2005.

Oliphant, T. E.: A Guide to NumPy, Trelgol Publishing USA. [online] Available from: http://ftp.sumy.volia.net/pub/FreeBSD/distfiles/numpybook.pdf (Accessed 14 August 2013), 2006.

Orlowski, N., Kraft, P., Pferdmenges, J. and Breuer, L.: Exploring water cycle dynamics by sampling multiple stable water isotope pools in a developed landscape in Germany, Hydrol. Earth Syst. Sci., 20(9), 3873–3894, doi:10.5194/hess-20-3873-2016, 2016.

Ortiz, C., Karltun, E., Stendahl, J., Gärdenäs, A. I. and Ågren, G. I.: Modelling soil carbon development in Swedish coniferous forest soils—An uncertainty analysis of parameters and model estimates using the GLUE method, Ecol. Model., 222(17), 3020–3032, doi:10.1016/j.ecolmodel.2011.05.034, 2011.

Over, M. W., Wollschläger, U., Osorio-Murillo, C. A. and Rubin, Y.: Bayesian inversion of Mualem-van Genuchten parameters in a multilayer soil profile: A data-driven, assumption-free likelihood function, Water Resour. Res., 51(2), 861–884, doi:10.1002/2014WR015252, 2015.

Papen, H. and Butterbach-Bahl, K.: A 3-year continuous record of nitrogen trace gas fluxes from untreated and limed soil of a N-saturated spruce and beech forest ecosystem in Germany: 1. N2O emissions, J. Geophys. Res. Atmospheres, 104(D15), 18487–18503, doi:10.1029/1999JD900293, 1999.

Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resour. Res., 42(5), W05302, doi:10.1029/2005WR004820, 2006.

Parsons, A. J.: The effects of season and management on the growth of grass swards, in The grass crop, pp. 129–177, Springer., 1988.

Parton, W. J., Stewart, J. W. and Cole, C. V.: Dynamics of C, N, P and S in grassland soils: a model, Biogeochemistry, 5(1), 109–131, 1988.

Parton, W. J., Hartman, M., Ojima, D. and Schimel, D.: DAYCENT and its land surface submodel: description and testing, Glob. Planet. Change, 19(1–4), 35–48, doi:10.1016/S0921-8181(98)00040-X, 1998.

Patil, A., Huard, D. and Fonnesbeck, C. J.: PyMC: Bayesian stochastic modelling in Python, J. Stat. Softw., 35(4), 1–81, 2010.

Perkel, J. M.: Programming: Pick up Python, Nature, 518(7537), 125–126, doi:10.1038/518125a, 2015.

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B. and Wagener, T.: Sensitivity analysis of environmental models: A systematic review with practical workflow, Environ. Model. Softw., 79, 214–232, doi:10.1016/j.envsoft.2016.02.008, 2016.

Potter, M. A. and Jong, K. A. D.: A cooperative coevolutionary approach to function optimization, in Parallel Problem Solving from Nature — PPSN III, edited by Y. Davidor, H.-P. Schwefel, and R. Männer, pp. 249–257, Springer Berlin Heidelberg. [online] Available from: http://link.springer.com/chapter/10.1007/3-540-58484-6_269 (Accessed 14 January 2015), 1994.

Rafique, R., Fienen, M. N., Parkin, T. B. and Anex, R. P.: Nitrous Oxide Emissions from Cropland: a Procedure for Calibrating the DayCent Biogeochemical Model Using Inverse Modelling, Water. Air. Soil Pollut., 224(9), 1–15, doi:10.1007/s11270-013-1677-z, 2013.

Raich, J. W. and Schlesinger, W. H.: The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate, Tellus B, 44(2), 81–99, doi:0.1034/j.1600-0889, 1992.

Raich, J. W. and Tufekciogul, A.: Vegetation and soil respiration: correlations and controls, Biogeochemistry, 48(1), 71–90, doi:10.1023/A:1006112000616, 2000.

Ravishankara, A. R., Daniel, J. S. and Portmann, R. W.: Nitrous oxide (N2O): the dominant ozone-depleting substance emitted in the 21st century, science, 326(5949), 123–125, doi:10.1126/science.1176985, 2009.

Reay, D. S., Davidson, E. A., Smith, K. A., Smith, P., Melillo, J. M., Dentener, F. and Crutzen, P. J.: Global agriculture and nitrous oxide emissions, Nat. Clim. Change, 2(6), 410–416, doi:10.1038/nclimate1458, 2012.

Rochette, P., Flanagan, L. B. and Gregorich, E. G.: Separating soil respiration into plant and soil components using analyses of the natural abundance of carbon-13, Soil Sci. Soc. Am. J., 63(5), 1207–1213, 1999.

Rosenbrock, H. H.: An automatic method for finding the greatest or least value of a function, Comput. J., 3(3), 175–184, 1960.

Saltelli, A., Tarantola, S. and Chan, K. P.-S.: A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output, Technometrics, 41(1), 39–56, doi:10.1080/00401706.1999.10485594, 1999.

Santos, C. A., Suzuki, K. and Watanabe, M.: Improvement in a genetic algorithm for optimization of runoff-erosion models, Annu. J. Hydraul. Eng., 44, 705–710, doi:10.2208/prohe.44.705, 2000.

Savage, K., Phillips, R. and Davidson, E.: High temporal frequency measurements of greenhouse gas emissions from soils, Biogeosciences, 11(10), 2709–2720, doi:10.5194/bg-11-2709-2014, 2014.

Schuëller, G. I. and Pradlwarter, H. J.: Benchmark study on reliability estimation in higher dimensions of structural systems – An overview, Struct. Saf., 29(3), 167–182, doi:10.1016/j.strusafe.2006.07.010, 2007.

Scott, R. L., Shuttleworth, W. J., Keefer, T. O. and Warrick, A. W.: Modeling multiyear observations of soil moisture recharge in the semiarid American Southwest, Water Resour. Res., 36(8), 2233–2247, doi:10.1029/2000WR900116, 2000.

Seifert, A.-G., Roth, V.-N., Dittmar, T., Gleixner, G., Breuer, L., Houska, T. and Marxsen, J.: Comparing molecular composition of dissolved organic matter in soil and stream water: Influence of land use and chemical characteristics, Sci. Total Environ., 571, 142–152, doi:10.1016/j.scitotenv.2016.07.033, 2016.

Senapati, N., Jansson, P.-E., Smith, P. and Chabbi, A.: Modelling heat, water and carbon fluxes in mown grassland under multi-objective and multi-criteria constraints, Environ. Model. Softw., 80, 201–224, doi:10.1016/j.envsoft.2016.02.025, 2016.

Shafii, M., Tolson, B. and Shawn Matott, L.: Addressing subjective decision-making inherent in GLUE-based multi-criteria rainfall–runoff model calibration, J. Hydrol., 523, 693–705, doi:10.1016/j.jhydrol.2015.01.051, 2015.

Siemens, J. and Kaupenjohann, M.: Contribution of dissolved organic nitrogen to N leaching from four German agricultural soils, J. Plant Nutr. Soil Sci., 165(6), 675–681, doi:10.1002/jpln.200290002, 2002.

Smith, K. A., Ball, T., Conen, F., Dobbie, K. E., Massheder, J. and Rey, A.: Exchange of greenhouse gases between soil and atmosphere: interactions of soil physical factors and biological processes, Eur. J. Soil Sci., 54(4), 779–791, doi:10.1046/j.1351-0754.2003.0567.x, 2003.

Smith, T., Sharma, A., Marshall, L., Mehrotra, R. and Sisson, S.: Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, Water Resour. Res., 46(12), W12551, doi:10.1029/2010WR009514, 2010.

Smith, T., Marshall, L. and Sharma, A.: Modeling residual hydrologic errors with Bayesian inference, J. Hydrol., 528, 29–37, doi:10.1016/j.jhydrol.2015.05.051, 2015.

Smith, T. J. and Marshall, L. A.: Bayesian methods in hydrologic modeling: A study of recent advancements in Markov chain Monte Carlo techniques, Water Resour. Res., 44(12), W00B05, doi:10.1029/2007WR006705, 2008.

Sone, C., Saito, K. and Futakuchi, K.: Comparison of three methods for estimating leaf area index of upland rice cultivars, Crop Sci., 49(4), 1438–1443, doi:10.2135/cropsci2008.09.0520, 2009.

Sorooshian, S., Duan, Q. and Gupta, V. K.: Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model, Water Resour. Res., 29(4), 1185–1194, doi:10.1029/92WR02617, 1993.

Soussana, J. F., Allard, V., Pilegaard, K., Ambus, P., Amman, C., Campbell, C., Ceschia, E., Clifton-Brown, J., Czóbel, S. Z. and Domingues, R.: Full accounting of the greenhouse gas (CO2, N2O, CH4) budget of nine European grassland sites, Agric. Ecosyst. Environ., 121(1), 121–134, doi:10.1016/j.agee.2006.12.022, 2007.

Stacey, A., Jancic, M. and Grundy, I.: Particle swarm optimization with mutation, in The 2003 Congress on Evolutionary Computation, 2003. CEC '03, vol. 2, p. 1425–1430 Vol.2., 2003.

Stehfest, E. and Bouwman, L.: N2O and NO emission from agricultural fields and soils under natural vegetation: summarizing available measurement data and modeling of global annual emissions, Nutr. Cycl. Agroecosystems, 74(3), 207–228, doi:10.1007/s10705-006-9000-7, 2006.

Storn, R. and Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces, J. Glob. Optim., 11(4), 341–359, 1997.

Subke, J.-A., Reichstein, M. and Tenhunen, J. D.: Explaining temporal variation in soil CO 2 efflux in a mature spruce forest in Southern Germany, Soil Biol. Biochem., 35(11), 1467–1483, doi:10.1016/S0038-0717(03)00241-4, 2003.

Suleau, M., Moureaux, C., Dufranne, D., Buysse, P., Bodson, B., Destain, J.-P., Heinesch, B., Debacq, A. and Aubinet, M.: Respiration of three Belgian crops: partitioning of total ecosystem respiration in its heterotrophic, above-and below-ground autotrophic components, Agric. For. Meteorol., 151(5), 633–643, doi:10.1016/j.agrformet.2011.01.012, 2011.

Tague, C. L. and Band, L. E.: RHESSys: Regional Hydro-Ecologic Simulation System—An object-oriented approach to spatially distributed modeling of carbon, water, and nutrient cycling, Earth Interact., 8(19), 1–42, doi:10.1175/1087-3562, 2004.

Thorburn, P. J., Biggs, J. S., Collins, K. and Probert, M. E.: Using the APSIM model to estimate nitrous oxide emissions from diverse Australian sugarcane production systems, Agric. Ecosyst. Environ., 136(3), 343–350, doi:10.1016/j.agee.2009.12.014, 2010.

Thornthwaite, C. W., 1899-1963, Mather, J. R. and 1923-: Instructions and tables for computing potential evapotranspiration and the water balance, [online] Available from: http://agris.fao.org/agris-search/search.do?recordID=US201300554032 (Accessed 7 December 2015), 1957.

Thyer, M., Kuczera, G. and Bates, B. C.: Probabilistic optimization for conceptual rainfall-runoff models: A comparison of the shuffled complex evolution and simulated annealing algorithms, Water Resour. Res., 35(3), 767–773, doi:10.1029/1998WR900058, 1999.

Vaché, K. B., McDonnell, J. J. and Bolte, J.: On the use of multiple criteria for a posteriori model rejection: Soft data to characterize model performance, Geophys. Res. Lett., 31(21), L21504, doi:10.1029/2004GL021577, 2004.

Van Genuchten, M. T.: A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, Soil Sci. Soc. Am. J., 44(5), 892–898, 1980.

Vrugt, J. A.: Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, Environ. Model. Softw., 75, 273–316, doi:10.1016/j.envsoft.2015.08.013, 2016.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W. and Sorooshian, S.: Effective and efficient algorithm for multiobjective optimization of hydrologic models, Water Resour. Res., 39(8), 1214, doi:10.1029/2002WR001746, 2003.

Vrugt, J. A., Ter Braak, C. J., Gupta, H. V. and Robinson, B. A.: Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?, Stoch. Environ. Res. Risk Assess., 23(7), 1011–1026, 2009.

Wagener, T. and Gupta, H. V.: Model identification for hydrological forecasting under uncertainty, Stoch. Environ. Res. Risk Assess., 19(6), 378–387, doi:10.1007/s00477-005-0006-5, 2005.

Wallach, D.: Evaluating crop models, Work. Dyn. Crop Models Elsevier Amst. Neth., 11–53, 2006.

Wang, G. and Chen, S.: A review on parameterization and uncertainty in modeling greenhouse gas emissions from soil, Geoderma, 170, 206–216, doi:10.1016/j.geoderma.2011.11.009, 2012.

Wang, S., Chen, M., Huang, D., Guo, X. and Wang, C.: Dream Effected Particle Swarm Optimization Algorithm, J. Inf. Comput. Sci., 11(15), 5631–5640, doi:10.12733/jics20104829, 2014.

Wang, Y.-P., Trudinger, C. M. and Enting, I. G.: A review of applications of model–data fusion to studies of terrestrial carbon fluxes at different scales, Agric. For. Meteorol., 149(11), 1829–1842, doi:10.1016/j.agrformet.2009.07.009, 2009.

Waring, R. H., Landsberg, J. J. and Williams, M.: Net primary production of forests: a constant fraction of gross primary production?, Tree Physiol., 18(2), 129–134, doi:10.1093/treephys/18.2.129, 1998.

Weise, T.: Global optimization algorithms-theory and application, Self-Publ. [online] Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.8184&rep=rep1&type=pdf (Accessed 19 March 2015), 2009.

Willmott, C. J.: On the validation of models, Phys. Geogr., 2(2), 184–194, 1981.

Windhorst, D., Kraft, P., Timbe, E., Frede, H.-G. and Breuer, L.: Stable water isotope tracing through hydrological models for disentangling runoff generation processes at the hillslope scale, Hydrol Earth Syst Sci, 18(10), 4113–4127, doi:10.5194/hess-18-4113-2014, 2014.

Wlotzka, M., Haas, E., Kraft, P., Heuveline, V., Klatt, S., Kraus, D., Butterbach-Bahl, K. and Breuer, L.: Dynamic Simulation of Land Management Effects on Soil N2O Emissions using a coupled Hydrology-Ecosystem Model, Prepr. Ser. Eng. Math. Comput. Lab, (03), 1–20, doi:10.11588/emclpp.2013.03.11824, 2013.

Wlotzka, M., Heuveline, V., Klatt, S., Haas, E., Kraus, D., Butterbach-Bahl, K., Kraft, P. and Breuer, L.: Simulation of Land Management Effects on Soil N2O Emissions using a Coupled Hydrology-Biogeochemistry Model on the Landscape Scale, in Handbook of Geomathematics, pp. 1–22, Springer., 2014.

Wohlfahrt, G., Bahn, M., Haslwanter, A., Newesely, C. and Cernusca, A.: Estimation of daytime ecosystem respiration to determine gross primary production of a mountain meadow, Agric. For. Meteorol., 130(1), 13–25, doi:10.1016/j.agrformet.2005.02.001, 2005a.

Wohlfahrt, G., Anfang, C., Bahn, M., Haslwanter, A., Newesely, C., Schmitt, M., Drösler, M., Pfadenhauer, J. and Cernusca, A.: Quantifying nighttime ecosystem respiration of a meadow using eddy covariance, chambers and modelling, Agric. For. Meteorol., 128(3), 141–162, doi:10.1016/j.agrformet.2004.11.003, 2005b.

Wolf, B., Zheng, X., Brüggemann, N., Chen, W., Dannenmann, M., Han, X., Sutton, M. A., Wu, H., Yao, Z. and Butterbach-Bahl, K.: Grazing-induced reduction of natural nitrous oxide release from continental steppe, Nature, 464(7290), 881–884, doi:10.1038/nature08931, 2010.

Xiang, S.-R., Doyle, A., Holden, P. A. and Schimel, J. P.: Drying and rewetting effects on C and N mineralization and microbial activity in surface and subsurface California grassland soils, Soil Biol. Biochem., 40(9), 2281–2289, doi:10.1016/j.soilbio.2008.05.004, 2008.

Yang, J., Reichert, P., Abbaspour, K. C., Xia, J. and Yang, H.: Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China, J. Hydrol., 358(1–2), 1–23, doi:10.1016/j.jhydrol.2008.05.012, 2008.

Zhang, W., Liu, C., Zheng, X., Zhou, Z., Cui, F., Zhu, B., Haas, E., Klatt, S., Butterbach-Bahl, K. and Kiese, R.: Comparison of the DNDC, LandscapeDNDC and IAP-N-GAS models for simulating nitrous oxide and nitric oxide emissions from the winter wheat–summer maize rotation system, Agric. Syst., 140, 1–10, doi:10.1016/j.agsy.2015.08.003, 2015.

Zhu, G. and Kwong, S.: Gbest-guided artificial bee colony algorithm for numerical function optimization, Appl. Math. Comput., 217(7), 3166–3173, doi:10.1016/j.amc.2010.08.049, 2010.

Acknowledgements

This dissertation is a group effort and would not have been possible without the help of many people:

At first, I would like to thank my supervisor Lutz for his open, straightforward and professional support throughout my time at the institute. He made this whole project easy-going and kept me motivated on track.

Special thanks deserves Sebi. He had the time consuming task of making a PhD student out of me. He showed me the power of coding and skills on Mountain bikes, I would not have imagined would be possible.

Further thanks deserves Philipp, who prevented on the one site numerous coding errors. On the other site, he managed to provide a safe harbor for all the data, which was generated throughout this project.

I owe many thanks to the “Schwingbach taskforce”, i.e. all the people gathered the data, that I used in this dissertation: Johannes Laufer, Björn Weeser, Konrad Bestian, Alice Aubert, Ina Plesca, Florian Lauer, Natalie Orlowski, Joscha Ufermann, Stefan Lübke, Felix Kruck, Patrick Widrinski, Gianna Dehler, Annika Diekl, Romina Gehler, Michael Herzog, Katrin Huber, Vanessa Jung, Melanie Oberhauser, Eva Holthof, Hartmut Holly, Alexander Konrad, Thosten Kühnel, Tim Grzelachowski, Katharina Schnaubelt, Ole Schnepel, Reiner Sigle, Michael Stein, Jaqueline Stenfert-Kroese, Nicole Werstein and Timm Zöltzer. As well as the friendly people from the lab: Heike Weller, Beate Lindenstruth, Nelly Weis, Günter Weber, Julia Valverde and Anja Schäfler-Schmid.

To all other colleagues and friends from our institute and the KIT in Garmisch-Partenkirchen: Thank you for the nice time and the amusing discussions during lunch, coffee and beer.

I thank my parents for their continuous support. They are the fundament of my life and enabled me to do the things that I want to do. Finally yet importantly, I thank Juliane for being the highlight in my life and all the adventures we started so far.

Declaration

I declare that I have completed this dissertation single-handedly without the unauthorized help of a second party and only with the assistance acknowledged therein. I have appropriately acknowledged and cited all text passages that are derived verbatim from or are based on the content of published work of others, and all information relating to verbal communications. I consent to the use of an anti-plagiarism software to check my thesis. I have abided by the principles of good scientific conduct laid down in the charter of the Justus Liebig University Giessen „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ in carrying out the investigations described in the dissertation.”

Gießen, 14th June 2017

_____

Tobias Houska