Reviewer #1

We thank the reviewer for its valuable comments which helped us improving the manuscript by re-analysing the validation results and the residual error distributions, amongst other things.

The paper is generally well written. I especially liked the methodology part. Although the paper is methodology oriented it remained accessible due to the clarity of writing in this section. However, this clarity is not reflected throughout the whole text and the manuscript lacks a clear focus. Mainly, while the manuscripts tries to increase model performance (addressing a changing measurement error variance) for four grassland it does not accomplishes this. The end of paragraph 4.2.2 summarizes the paper succinctly: “Overall, while the HE2 inversion framework is arguably more conceptually sound, we found that it does not permit to fully remove heteroscedasticity from the residuals while simultaneously leading to a poorer modeling performance in terms of fitting the large observed values.”!

I would like to see the authors address this issue. WHY doesn’t it work? What in the model structure might cause this behaviour? Can you address this and make it work? As it stands you have not provided any potential solutions to the measurement error variance problem. Although this shouldn’t be an issue if potential solutions are offered or discussed (literature review using studies in other ecosystems?) this isn’t the case.

We agree that the former version of the paper lacked of a clear focus. Actually, the former manuscript had two objectives, as stated previously in the introduction: 1) to compare the inversion of eddy covariance data with different treatments of eddy covariance uncertainties and 2) to perform a cross-site comparison. We now focus the paper on the first point, although we maintain the cross-site comparison as this remains an interesting issue for ecophysiologists.

The inversions conducted in the framework of this paper did increase the CARAIB model performances compared to using default model parameter values (see Table 2, \textsuperscript{3r}d column) as done in previous studies. Furthermore, a particular attention has to be given to having rigorous posterior distributions from a statistical point of view: that's why the heteroscedastic residual error models are worth being tested, because eddy covariance data show heteroscedasticity.

The revised version now more clearly concludes on the outcomes of the comparison of the different ways of treating eddy covariance residual errors in the inversions. Several new analyses and discussions were also included. These are detailed below:

- We computed performance indicators using the validation dataset not only for the HO2 scenarios but for all scenarios. The corresponding results are quite interesting: heteroscedastic inversions actually outperform homoscedastic inversions in validation, which was not the case in calibration. Posterior parameter sets obtained from the heteroscedastic inversions thus appear to be more robust. This new finding changed our conclusion regarding the most appropriate way of treating eddy covariance residual errors, both from a theoretical and practical vegetation modeler’s point of view. Please see in the revised version a new table (Table 6) summarizing the validation for all inversion scenarios and a new figure of the measured and modelled GPP at Monte-Bondone for the scenarios HO2 (homoscedastic) and HE2 (heteroscedastic) (Figure 6). For the validation period, it is observed that the uncertainty bands due to parameter uncertainty (dark gray area) derived from using the HE2 model bracket more closely the measured signal
(green line) compared to the uncertainty bands due to parameter uncertainty derived from using the HO2 model. Lastly, it is worth noting that the HE2 model allows for tight total predictive uncertainty bands (light gray area) around zero when nearly-null GPP is modelled. This is much more consistent than the overly large total predictive uncertainty bands around zero derived with the HO2 model, which unrealistically comprises largely negative GPP values.

- We carefully looked at the standardized residuals associated with every site (not only Monte-Bondone as presented in the former version of the paper) and observed that the use of the HE2 model did remove residual heteroscedasticity to some extent, although not fully. Even if we were aware of this fact before, the first version of the manuscript was not enough balanced on that point and may have suggested that the HE2 inversions did not succeed in removing heteroscedasticity at all, which is not the case. The standardized residuals as in Figure 5 of the revised manuscript are presented below for all the sites:

- As can be seen in Figures 3a, 4 and 5, a fraction of the residual error heteroscedasticity is caused by the inability of the CARAIB model to simulate the highest observed GPP values. Because of this model inadequacy, larger residual errors are more frequently observed for large observed GPP values. This model bias seems too complex to be handled within the likelihood
function (see, e.g., Reichert and Schuwirth, 2012, for research in this direction). As for any model bias this calls for model improvement.

Considering all these remarks, we updated the paper in order to (1) focus on the refined scope and (2) give better recommendations for future research. Here are the main modifications that we have made to the manuscript:

- The following sentence in section 3.1.2 was modified (L360-362): “In the remainder of this document, results are mainly detailed for this inversion scenario, since it generally led to the lowest data misfit statistics in calibration.” instead of “In the remainder of this document, results are mainly detailed for this inversion scenario, because it generally led to the lowest data misfits while inversions HE1 and HE2 did not fully removed heteroscedasticity (see further).”
- It was detailed in L434 that the reported results concern calibration only: “Overall, in calibration, modelled signals...
- Figure 5 was better interpreted: the following line was added in section 3.2.3 (L436): “However, the total predictive uncertainty range derived from the HE2 inversions was more consistent, as, e.g., it avoids unrealistic negative values of GPP.”
- The following sentence was modified in section 3.2.3 (L441-443): “Heteroscedasticity of the GPP residual errors was fairly reduced but not fully removed by using the HE1 and HE2 heteroscedastic residual error models. Indeed, the standardised residuals still showed some small but complex heteroscedastic patterns.”
- A table (Table 6) summarizing validation results for all inversion scenarios and sites was added. A figure (Figure 6) highlighting the respective calibration and validations results of the HO2 and HE2 inversions has been added. The text of section 3.3 (Validation) has been modified as follows (L456-470): “Parameter values from the posterior distributions were tested for validation using eddy covariance data over different periods (validation datasets). Figure 6 shows measured and modelled GPP values over the periods of calibration and validation in Monte-Bondone. Not surprisingly, worse agreements between measured and modelled data are observed as compared to the calibration period. However, it is observed that the modelled GPP in validation in the HE2 inversions follows better the measured signal than in the HO2 inversions. Strikingly, in all the sites, the posterior parameter distributions derived from using the HE1 and HE2 heteroscedastic models are found to induce a better model performance in validation compared to the posterior distributions associated with the use of the homoscedastic models (Table 6). The difference between calibration and validation appeared thus smaller when using most likely parameter values from heteroscedastic inversions as compared to homoscedastic inversions. Among the different grassland sites, a similar performance pattern as for the calibration experiment is observed. Indeed, the Laqueuille site shows for each type of measurement data the worst performance statistics whereas the Monte-Bondone site overall presents the best fits to the data (Table 6).”
- Some sentences about the model limitation were added in “4.1 Measured and modelled signals” (L514-517): “Another modelling limitation is that model parameters are assumed as constant along the season, although plants traits are known to evolve throughout the season and plants acclimate to specific climate conditions. As a result, the effect of similar climatic conditions does not necessary result in similar eddy covariance measurements.”
- The section “4.2.1 Homoscedastic and heteroscedastic eddy covariance residual errors” in the
discussion was largely modified. The following sentences were added/modified (starting at L538): “However, in validation, the posterior parameter distributions derived from using the heteroscedastic residual error models outperform their counterparts derived from using the homoscedastic residual error models. This important finding reveals that despite inducing larger RMSE values in calibration, the use of a heteroscedastic residual error model leads to a more robust parameter estimation” (...) “It is also worth noting that a substantial fraction of the large residual errors is caused by the tendency of the CARAIB model of underestimating the observed GPP summer peaks. As discussed above, this is related to a slower temporal resolution of the model compared to that of the measured data. To overcome this model inadequacy, further model modifications are necessary to increase the time resolution of the model. Another model improvement would be to simulate varying model parameter values as a function of the time of the year, since plant traits are actually evolving along the seasons. However, this would come at the cost of a large increase in model complexity.”

- A sentence was modified in section 4.2.2 for stressing one advantage of the scenarios HO2 and HE2 (L561): “The benefit of these values is that they inform about...” instead of “These values inform about...”

- The following sentence in section 4.2.2 was deleted: “Overall, while the HE2 inversion framework is arguably more conceptually sound, we found that it does not permit to fully remove heteroscedasticity from the residuals (Figure 5) while simultaneously leading to a poorer modelling performance in terms of fitting the large observed values (such as the summer GPP).”

- The following sentence was modified in section 4.3 to clarify the effect of some model limitations on parameter values (L587-590): “As the model cannot simulate the high GPP values that are observed in the eddy covariance data, the Bayesian algorithm could have compensated by sampling high values of g1 that increase stomatal conductance.” instead of “As the model cannot simulate the fast dynamics of the carbon fluxes that are observed in the eddy covariance data, the Bayesian algorithm could have compensated by sampling high values of g1.”

- The abstract, objectives and conclusions were modified according to the above-mentioned changes in the interpretation of results.

Delete paragraph 4.4 and anything which leads to it. If you do not make an effort to optimize for a global parameter set there is little value in discussing the between site differences. It is well established that if your model lacks the specificity to address between site differences different parameterizations are necessary.

The reviewer #2 also suggested to shorten section 4.4 and to merge it with the previous section. We summarized this section in ~10 lines (L591-600), and merged it with the previous one. The text in section 4.3 was also shortened while keeping its key points. Abstract, objectives and especially conclusions were also modified accordingly. Please look at the new section 4.3 (starting at L566).

However, even if we focus the paper on the ways on treating eddy covariance data errors in the inversion, we want to keep the section 4.3 in the paper (as a second objective of the paper). We still believe that there is an interest in discussing the absolute values of the parameters such as SLA and g1.
in these four sites where a number of similar inversion studies have been conducted.

*Paragraph 2.3.2 should go in the introductions, as it sets up the issues of the uncertainty in eddy covariance measurements better than what is currently in the introduction.*

The section 2.3.2 was merged in the introduction and some statements were modified (L69-85). This new paragraph about eddy covariance data uncertainties in the introduction better sets up the current research needs and the scope of the paper, that is on the way of treating random eddy covariance data errors in the inversion. As section 2.3.1 moved to a new section (2.1 Experimental sites and data), the section 2.3 « Eddy covariance data » was deleted.

Merge 2.2 and 2.3.1 and put this before the model description (I assume this was originally the case as the order of tables 1 and 2 is reversed - first addressing table 2 then 1)

Sections 2.2 and 2.3.1 were moved to a new section entitled “2.1 Experimental sites and data”. The text was slightly modified and some statements from 2.3.2 were also moved to this section.

Other changes:
- As we moved some part of the “Materials & Methods” to the introduction, some part of the introduction that were too wordy were shortened, i.e., the paragraph about the representativeness of the data used for inversion (L50-55).
- The objectives of the paper presented in the introduction were modified to stress the scope of the paper that is on the comparison of different ways of treating eddy covariance data errors (L86-95).
- Abstract and conclusions were modified to reflect the refined scope of the paper.
Reviewer #2

We thank the reviewer for his useful comments and positive assessment about our work. We addressed all the points raised by the reviewer. Please see below.

1. I think the most interesting part of this manuscript is the discussion of the heteroscedastic measurement errors (Section 4.2). More discussions on why the current linear heteroscedastic model doesn’t work well would be plausible. I think your winter-summer discussion is a good start. (page 1815, line 10-14).

We are totally in line with the point that the most interesting and innovative part is about the comparison of the way eddy covariance data uncertainties are treated, including the homoscedastic versus heteroscedastic residual error models. As pointed out by the first reviewer, the former version of the paper did lack of clear focus. The current revised version now deeper discusses the differences between the inversion scenarios and provides a more comprehensive analysis of the validation experiment.

Furthermore, we would like to highlight that it should not be concluded from our study that using a heteroscedastic residual error model is not worth it. Since eddy covariance measurement errors show heteroscedasticity, using a heteroscedastic residual error model is arguably more statistically sound. More important, a closer look at our results revealed that parameter estimation using a heteroscedastic error model rather than a homoscedastic error model actually leads to a better modelling performance in validation. This new finding provides support for the use of a heteroscedastic error model. This is detailed in the current version of the paper.

Below are the modifications that we made to address this issue:

• The section “4.2.1 Homoscedastic and heteroscedastic eddy covariance residual errors” in the discussion was largely modified. Please look at the answer of the 1st comment of reviewer #1 or in the manuscript.
• A sentence was modified in section 4.2.2 for stressing one advantage of the scenarios HO2 and HE2 (L561): “The benefit of these values is that they inform about...” instead of “These values inform about...”
• The following sentence in section 4.2.2 was deleted: “Overall, while the HE2 inversion framework is arguably more conceptually sound, we found that it does not permit to fully remove heteroscedasticity from the residuals (Figure 5) while simultaneously leading to a poorer modelling performance in terms of fitting the large observed values (such as the summer GPP),”
• The abstract, objectives and conclusions were modified according to the above-mentioned changes in the interpretation of results.

2. A general outline of how the inversion works at the four sites is suggested to include at the start of Section 2.4, which should help on a clear technical road map of the paper.

A paragraph was modified at the end of former section 2.4.4 (now section 2.3.4) that summarizes the four inversion scenarios used in this study. We believe that the whole section 2.3.4 provides enough
The new paragraph reads (L311-320): “This treatment of multi-objective Bayesian inference is in line with the work of Reichert and Schuwirth (2012), who further considered different statistical models for model and observation errors. Overall, this resulted in four different ways of treating the eddy covariance data uncertainties: fixed homoscedastic (HO1) and heteroscedastic (HE1) error models, and jointly inferred homoscedastic (HO2) and heteroscedastic (HE2) error models. Using the HO1 and HE1 models led to a total of 10 inferred parameters, whereas using the HO2 and HE2 models resulted into a total of 13 and 16 inferred parameters, respectively. Table 2 lists the marginal prior distributions used for all sampled parameters. The upper and lower bounds of the prior parameter distributions were set using boundary values that correspond at least to the lower and upper physically-possible bounds of the parameters, or to narrower bounds using expert-knowledge.”

3. Section 4.4 could be merged into section 4.3 but should be more concise. For example, the length of discussion on the difference of the parameter values across sites could be reduced. Discussions on how you learn from your multi-site bayesian inversion study to design a common parameter set across sites and the advantages and disadvantages of the two options you gave at the end of section 4.4 should be extended. Otherwise they should be removed as they looks too thin. Your choice.

Kuppel et al., 2012 could be a potential reference.


Reviewer #1 also suggested shortening section 4.4. We decide to summarize section 4.4 in ~10 lines and merge it in section 4.3. Please look at L591-600 in the revised manuscript. We thank the reviewer for pointing out the Kuppel et al. (2012) study which is really relevant to our discussion about multi-site versus single site inversion.

The section 4.3 was also shortened while keeping its key points. Abstract, objectives and conclusions were also modified accordingly. Overall, the paper pays now less attention to the cross-site comparison and focus more on the treatment of eddy covariance data uncertainties.

4. Typo error: H02 should be HO2 in the title of Table 5.
OK

5. The unit of ET should be mm day-1 throughout the manuscript.
OK

Other changes:
- As we moved some part of the “Materials & Methods” to the introduction, some part of the introduction that were too wordy were shortened, i.e., the paragraph about the representativeness of the data used for inversion (L50-55).
- The objectives of the paper presented in the introduction were modified to stress the scope of
the paper that is on the comparison of different ways of treating eddy covariance data errors (L86-95).

- Abstract and conclusions were modified to reflect the refined scope of the paper.
Bayesian inversions of a dynamic vegetation model in four European grassland sites

Julien Minet¹, Eric Laloy², Bernard Tychon¹, and Louis François³

¹Université de Liège, Arlon Campus Environnement, Avenue de Longwy 185, B-6700 Arlon, Belgium
²Belgian Nuclear Research Centre (SCK-CEN), Boerentang 200, B-2400 Mol, Belgium
³Université de Liège, UMCCB, Allée du six août 17, B-4000 Liège, Belgium

Correspondence to: Julien Minet (julien.minet@ulg.ac.be)

Abstract. Eddy covariance data from four European grassland sites are used to probabilistically invert the CARAIB dynamic vegetation model (DVM) with ten unknown parameters, using the DREAM[ZS] Markov chain Monte Carlo (MCMC) sampler. We especially compare model inversions considering both homoscedastic and heteroscedastic eddy covariance residual errors, with variances either fixed a priori or jointly inferred with the model parameters. Agreements between measured and simulated data during calibration are comparable with previous studies, with root-mean-square error (RMSE) of simulated daily gross primary productivity (GPP), ecosystem respiration (RECO) and evapotranspiration (ET) ranging from 1.73 to 2.19 gC m⁻² day⁻¹, 1.04 to 1.56 gC m⁻² day⁻¹, and 0.50 to 1.28 mm day⁻¹, respectively. For the calibration period, using a homoscedastic eddy covariance residual error model resulted in a better agreement between measured and modelled data than using a heteroscedastic residual error model. However, a model validation experiment showed that CARAIB models calibrated considering heteroscedastic residual errors perform better. Posterior parameter distributions derived from using a heteroscedastic model of the residuals thus appear to be more robust. This even though the classical linear heteroscedastic error model assumed herein did not fully removed heteroscedasticity of the GPP residuals. Despite the fact that the calibrated model is generally capable of fitting the data within measurement errors, systematic bias in the model simulations are observed. These are likely due to model inadequacies such as shortcomings in the photosynthesis modelling. Besides the residual error treatment, differences between model parameter posterior distributions among the four grassland sites are also investigated. It is shown that the marginal distributions of the specific leaf area and characteristic mortality time parameters can be explained by site-specific ecophysiological characteristics.
1 Introduction

Covering about 38% of the European agricultural area and 8% of the land surface ([FAO] 2011), grassland is an important land cover class in Europe which shows a wide range of different ecological characteristics. By stocking carbon, temperate grassland might play an important role in climate change mitigation in Europe ([Soussana et al.] 2004) and at the world-scale ([O’Mara] 2012). Large uncertainties however remain in the estimation of the (source or sink) carbon fluxes since those largely depend on farming management options.

In environmental modelling, grassland growth models have received less attention than the long-standing and highly-developed crop models. Since grasslands are agro-ecosystems that can be considered either as agricultural or semi-natural lands, grassland models were designed for two main purposes: the simulation of forage and dairy or meat production, and the simulation of the carbon fluxes at the land-atmosphere interface. Several crop models were adapted for grassland growth modelling (e.g., STICS ([Ruget] 2009 [Dumont et al.] 2014), EPIC ([Williams et al.] 2008)) especially when the management of the grassland remains similar to crop management, i.e., when the grassland is a temporary forage production that is cut rather than grazed by animals. Some other models were specifically developed for grasslands (e.g., SPACSYS ([Wu et al.] 2007)), sometimes coupled with animal production models (e.g., PASIM ([Graux et al.] 2013)), whereas grassland models were also developed from dynamic vegetation models (DVM) such as LPJmL ([Bondeau et al.] 2007), adapted from the LPJ model ([Sitch et al.] 2003). Being process-based models, DVM are well suited for large-scale spatial simulations and can account for a wide range of current and projected climatic conditions.

To be used for simulation-based decision making, a DVM must be properly parametrised. Model parameter values can be derived from (1) laboratory experiments as, e.g., the stomatal conductance described by the Ball-Berry model ([Ball et al.] 1987), (2) in-situ field measurements, and (3) model inversion using calibration data measurements or (4) spatialized databases (e.g., from remote sensing). Model inversion (also referred to as calibration) consists of automatically finding those model parameters that allow the model to adequately reproduce the available observed data. The collection of representative and high-quality data is thus of paramount importance for inversion, as DVMs require an adequate parametrization that is sufficiently representative of the range of conditions over the spatial extent of the simulation. Typically, DVMs use different set of parameters that are assigned to specific vegetation classes that grow together over the same area or in geographically distinct biomes. Dynamic vegetation model inversion needs a sufficient number of sites with varying ecophysiological conditions that are supposed to be representative of the considered vegetation classes or biomes, but still well-delimited ([Knorr and Kattge] 2005). Model inversion using continuous, gridded data (e.g., from remote sensing ([Patenaude et al.] 2008)) could also help in determining optimal parameters for large areas, but computation time can be a limiting factor for such application.
Given the high number of eddy covariance experimental sites across the world, eddy covariance measurements are particularly appealing for inversion of DVM models (Friend et al., 2007). Furthermore, the long-standing rise in computational resources not only increased modelling capabilities in terms of temporal and spatial resolution, but also opened new avenues for quantifying the uncertainty associated with the estimated model parameters and its effect on model simulations. In particular, the Bayesian framework for inverse modelling is increasingly used in the DVM community (e.g., Hartig et al., 2012). Bayesian methods such as Markov chain Monte Carlo (MCMC) sampling aim to derive a representative set of all parameter combinations that are consistent with the observed data and available prior information. This set of parameters is referred to as the posterior distribution.

Nevertheless, eddy covariance data are known to be associated with relatively large measurement uncertainties, implying both systematic and random errors (see Aubinet et al. (2012), chapter 7, for a comprehensive description of all sources of eddy covariance uncertainties). As eddy covariance data are the result of a long process chain, they can be affected by instrumental measurement error (e.g., calibration and design errors), sampling errors due to the variability of the fluxes in time and space and data treatment error (e.g., due to the gap-filling of missing data). Uncertainties in eddy covariance data is also strongly dependent on the time resolution of the fluxes, tending to diminish with time aggregation (Richardson and Hollinger, 2005). It is crucial to account for these random data uncertainties in the inversion since an improper statistical treatment can cause the parameter posterior distribution to be strongly biased (e.g., Fox et al., 2009). Quantifying random eddy covariance data errors is not straight-forward (Hollinger and Richardson, 2005; Lasslop et al., 2008), but these errors are typically characterized by a variance that is proportional to the magnitude of the data, i.e., they show heteroscedasticity (e.g., Lasslop et al., 2008). Therefore, it has been suggested (Richardson et al., 2008) that the measurement error variance can be modelled as a linear function of the magnitude of the flux with a non-null intercept, as random errors are non-null even when the flux equals zero. Yet, while the random error can be taken into account in the inversion, systematic measurement errors can only be removed by instrument calibration.

In this study, data from eddy covariance stations over four grassland sites are inverted for the CARAIB dynamic vegetation model parameters within a Bayesian framework. This is both the first automatic calibration of the CARAIB model and its first application to managed grassland modelling, which required adaptations of the model to grass cutting and grazing. The main objective is to compare the modelling of the carbon and water fluxes over the four grassland sites using four different ways of treating the eddy covariance data errors during the inversion. Both homoscedastic and heteroscedastic residual error models are considered, either fixed beforehand or sampled along with the model parameters. A second objective is then to compare the site-specific posterior parameter distributions obtained for the four grasslands, given their climatic, ecological and management characteristics.
2 Materials and Methods

2.1 Experimental sites and data

In this study, we focus on four long-term experimental sites (see table 1) that are semi-natural permanent grasslands: Grillenburg, Germany, (Prescher et al., 2010); Oensingen (intensive), Switzerland (Ammann et al., 2007); Monte-Bondone, Italy, (Wohlfahrt et al., 2008) and Laqueuille (extensive), France, (Klumpp et al., 2011). The four sites pertain to the global FLUXNET network and, as such, a large number of studies were conducted using eddy covariance data from these sites. The FLUXNET website (http://fluxnet.ornl.gov/) provides lists of references per site.

The four sites are located in western and central Europe and experience different climate, altitude, soil and management conditions. They can be classified according to the De Martonne-Gottman aridity index, which is inversely related with the site aridity. Oensingen is the most intensively managed site and the only one that is fertilized (about 200 kg N ha$^{-1}$ yr$^{-1}$). The other three sites are extensively managed, with no organic nor mineral fertilization. The last two sites are mid-mountainous grassland, while the first two sites are situated at a lower altitude. Only the grassland in Laqueuille is grazed by animals during the growing season, while the other three are hay meadows that are cut once or several times a year. Note that, although grass cutting should have occurred on the 13 June 2005 in Grillenburg according to the given management data, it was not observed in the measured eddy covariance fluxes because of gap-filling of missing data. As a result, this cut was neglected in the modelling.

The four grasslands are equipped with eddy covariance stations for measuring ecosystem fluxes. Data of flux measurement and field datasets were made available through a coordinated task of the FACCE/MACSUR knowledge hub, which aims at performing an intercomparison of grassland models (Ma et al., 2014) by running several grassland model with the same field datasets collected under various climatic and management conditions. Field datasets hold the necessary information for feeding the grassland model: hourly meteorological records of climatic variables, soil physical parameters, management information such as cutting dates or grazing charges, and initial conditions. Daily eddy covariance data included net ecosystem exchange (NEE) [gC m$^{-2}$ day$^{-1}$], gross primary productivity (GPP) [gC m$^{-2}$ day$^{-1}$], ecosystem respiration (RECO) [gC m$^{-2}$ day$^{-1}$] and evapotranspiration (ET) [mm day$^{-1}$]. It is worth noting that only the NEE and ET are directly measured by the eddy covariance station (i.e., fluxes of CO$_2$ and H$_2$O, respectively) and that GPP and RECO are derived from these measurements.

In this study, only GPP, RECO and ET measurements were used in the inverse modelling. Adding NEE measurements would be useless as they are directly dependent on GPP and RECO. The GPP and RECO were used since they are directly linked with the photosynthesis and respiration processes, respectively, while the influence of these two processes is mixed in the NEE measurements. Other combinations including the NEE were first tested but it resulted in poorer agreements between
measured and modelled data. The full data range including gap-filled data was inverted, since these
data are gap-filled according specific protocols that are standards in the eddy covariance community.

2.2 The CARAIB model

2.2.1 Description of the model

CARAIB is a physically-based dynamic vegetation model that was developed for the simulation of
the carbon cycle at the global scale (Warnant et al., 1994; Nemry et al., 1996; Otto et al., 2002). It
calculates the carbon fluxes through the soil-vegetation-atmosphere continuum by simulating eco-
physiological processes: photosynthesis, carbon allocation to plant pools and autotrophic and heter-
otrophic respiration. The CARAIB model has been used in numerous paleoclimatology, vegetation
and crop modelling studies. The reader is referred to the aforementioned references for full model
description.

For C3 plants, photosynthesis is computed according the model of Farquhar et al. (1980). The
stomatal conductance governing the flux of CO$_2$ through the stomata is described at the leaf scale
with the Ball-Berry approach (Ball et al., 1987), using the model of Leuning (1995) with further
adaptations from Van Wijk et al. (2000) for accounting for soil water stress affecting the stomatal
conductance. Photosynthesis and respiration processes are computed at a two-hour time step on a
half-day basis and the model assumes a symmetry with respect to solar noon time, that is, computa-
tion of these processes are made for half the day and further aggregated at a daily time step. Other
processes, e.g., related to soil hydrology or carbon allocation, are computed on a daily basis.

In this study, a single plant functional type (PFT) is considered (BAG 22 as defined in Laurent
et al., 2004, 2008) corresponding to the flora that can be encountered in European grasslands, i.e.,
species of Poaceae and Asteraceae. The model was adapted for simulating the grassland sites by
adding management functions for grass cutting and grazing. Grass cutting is modelled by the removal
of a part of the plant carbon mass so that the model matches given values of leaf area index after
cutting. Grazing is modelled such as a given fraction of the plant carbon mass is removed every day
according to the grazing charge. The dates of the grass cutting and the duration of the grazing periods
were known and fixed in the simulations. Daily meteorological data recorded at the experimental
sites were used in the model, i.e., minimal and maximal temperature, precipitation, solar radiation,
relative air humidity and wind velocity. Although they can affect vegetation modelling (Gottschalk
et al., 2007; Rivington et al., 2006; Zhao et al., 2012), uncertainties in the meteorological data were
not considered in this study.

Thirty-three parameters per PFT are set in CARAIB. These parameters govern photosynthesis,
plant physiology process (e.g., specific leaf area, carbon-to-nitrogen ratio), allocation of carbon
and residence times in the different pools of carbon including plants and soil pools, land surface-
atmosphere interactions (albedo, roughness length) and tolerance to extreme conditions (thresholds
and response times). During the model development, parameter values in CARAIB were mainly found in the literature (Warnant, 1999) and further compared with observed values (remote sensing, field data and paleorecords). So far, no model inversions were performed with the CARAIB model.

2.2.2 Choice of parameters

In this study, ten model parameters were sampled (Table 2). They were chosen according to their presupposed importance, that is, the model sensitivity to these parameters, and because some parameters values were already known in the measured data from the experimental sites. Default values that were defined during the model development and used in previous researches are given in Table 2. These parameters governs the main processes of the model, namely, the photosynthesis, the respiration and carbon transfer between carbon pools:

- The slope $g_1$ and the intercept $g_0$ [$\mu$mol.m$^{-2}$.s$^{-1}$] of the stomatal conductance as described in Leuning (1995) are directly related to the photosynthesis since they govern the stomatal conductance. They are thus related to the gross primary productivity (GPP) and evapotranspiration (ET) with respect to the meteorological conditions. While most of ecological models, including CARAIB, use an empirical approach for stomatal conductance derived from the Ball-Berry model, Medlyn et al. (2011) recently reconcile the empirical approach with the theoretical background based on the optimal stomatal behaviour (Farquhar et al., 1980), which states that there is a trade-off for stomata between maximizing carbon gain (photosynthesis) and minimizing water loss (transpiration). These new developments in the theoretical understanding of the empirical relationship push forward the necessity to measure or calibrate the stomatal conductance parameters under different environmental conditions. Although single values of these parameters are used for regional or global modelling of C3 plants photosynthesis (e.g., Sitch et al., 2008), it is actually known that stomatal conductance parameters should vary through time and space according to the environmental conditions and plant species.

- The specific leaf area (SLA) [m$^2$gC$^{-1}$] is defined in CARAIB as the leaf area per unit of carbon mass of the plants. It is used in the model to convert the assimilated mass of carbon into leaf area index. Besides its role in the model, SLA is often studied as a plant trait that is used for predicting the plant resource use strategy or for clustering plants species into functional groups. Maximizing the photosynthesis while minimizing leaf respiration, high SLA leaves (thin leaves) are productive, but also more vulnerable and short-lived (Wilson et al., 1999). They are thus better adapted to resource-rich environment, where leaves can be quickly reconstructed (Poorter and De Jong, 1999). At the other side, low SLA leaves (thick leaves) are often encountered in drought-adapted (Marcelis et al., 1998) or shade-tolerant species (Evans and Poorter, 2001) and for the lower, self-shaded leaves of a plant. SLA is also known to vary along the season and according to the leaf age (Wilson et al., 1999). Nevertheless, the concept
of SLA is sometimes problematic for some plant species with complex plant geometry (Vile et al., 2005), e.g., highly folded leaves, or with a non-negligible part of the photosynthetic tissues standing on the stem, as encountered among the Poaceae species. In these simulations, SLA is defined for the PFT that is supposed to represent European grasslands and therefore, it should be actually considered as an effective parameter among the grassland species and for the whole plant body.

- The characteristic mortality time [year] of the plant in normal $\tau$ and in stress conditions $\tau_s$ are, respectively, the characteristic time for the renewal of the plant ($\tau$) and the time it takes to the plant to die in stress conditions ($\tau_s$). The stress conditions occur when temperatures reach either low or high extreme values, for soil water content below a certain threshold or for low irradiance values. The default values were 0.667 year for $\tau$, meaning a renewal of the plant by 8 months, and 0.083 year for $\tau_s$, meaning a characteristic mortality time in stress conditions of one month.

- Two carbon-to-nitrogen ratio are defined for the photosynthetic active carbon pool of the plant (C/N1) and for the remainder of the plant (C/N2). The nitrogen content of the leaves play a crucial role in the photosynthesis and increasing nitrogen content (decreasing C/N) fosters photosynthetic activity. Low C/N ratio in plant usually comes together with high nitrogen content in soils, that is, a resource-rich environment.

- Three parameters govern the rates of the soil heterotrophic respiration: $\gamma_1$ for the respiration of the “green litter”, $\gamma_2$ for the respiration of the "not-green litter” and $\gamma_3$ for the respiration of the soil organic carbon.

### Table 1. Grassland sites and periods of simulations

<table>
<thead>
<tr>
<th>Coordinates</th>
<th>Altitude</th>
<th>Management</th>
<th>Fertilisation</th>
<th>De Martonne-Gottman index</th>
<th>Calibration years</th>
<th>Validation years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grillenburg, DE</td>
<td>13.50 °E 50.95 °N</td>
<td>380 m</td>
<td>cutting (1-3 yr⁻¹)</td>
<td>no</td>
<td>32</td>
<td>2004-2006</td>
</tr>
<tr>
<td>Oensingen, CH</td>
<td>7.73 °E 47.28 °N</td>
<td>450 m</td>
<td>cutting (3-5 yr⁻¹)</td>
<td>yes</td>
<td>38</td>
<td>2002-2005</td>
</tr>
<tr>
<td>Monte-Bondone, IT</td>
<td>11.03 °E 46.00 °N</td>
<td>1500 m</td>
<td>cutting (1 yr⁻¹)</td>
<td>no</td>
<td>35</td>
<td>2003-2005</td>
</tr>
<tr>
<td>Laqueuille, FR</td>
<td>2.73 °E 45.63 °N</td>
<td>1040 m</td>
<td>grazing</td>
<td>no</td>
<td>41</td>
<td>2004-2007</td>
</tr>
</tbody>
</table>

### 2.3 Probabilistic inversion methodology

#### 2.3.1 Inverse problem

To acknowledge that measurements and modelling errors are inevitable, the inverse problem is commonly represented by the stochastic relationship

$$F(z) = d + e,$$  \hspace{1cm} (1)
where $F$ is a deterministic, error-free forward model that expresses the relation between the uncertain parameters $z$ and the measurement data $d$, and the noise term $e$ lumps measurement and model errors.

Inversions were performed within a Bayesian framework, which treats the unknown model parameters $z$ as random variables with posterior probability density function (pdf) $p(z|d)$ given by

$$p(z|d) = \frac{p(z)p(d|z)}{p(d)} \propto p(z) L(z|d),$$

where $p(z)$ denotes the prior distribution of $z$ and $L(z|d) \equiv p(d|z)$ signifies the likelihood function of $z$. The normalization factor $p(d) = \int p(z)p(d|z)dz$ is obtained from numerical integration over the parameter space so that $p(z|d)$ scales to unity. The quantity $p(d)$ is generally difficult to estimate in practice but is not required for parameter inference. In the remainder of this study, we will focus on the unnormalized posterior $p(z|d) \propto p(z) L(z|d)$. For numerical stability, it is often preferable to work with the log-likelihood function, $\ell(z|d)$, instead of $L(z|d)$. If we assume the error $e$ to be normally distributed, uncorrelated and with unknown constant variance, $\sigma^2$, the log-likelihood function can be written as

$$\ell(z|d) = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\sigma^2) - \frac{1}{2 \sigma^2} \sum_{i=1}^{N} [d_i - F_i(z)]^2,$$

where $\sigma$ can be fixed beforehand or sampled jointly with the other model parameters $z$.

The homoscedasticity (i.e., constant variance) assumption for $e$ may be excessively strong in many cases. Considering the residual errors, $e$, to be heteroscedastic, Eq. (3) becomes

$$\ell(z|d) = -\frac{N}{2} \log(2\pi) - \sum_{i=1}^{N} \log(\sigma_i) - \frac{1}{2} \sum_{i=1}^{N} \frac{[d_i - F_i(z)]^2}{\sigma_i^2},$$

where the $\sigma_i$ are the individual residual error standard deviations, that can be gathered into a vector $\sigma$. Here also, $\sigma$ can either be fixed beforehand or sampled along with $z$ (see further).

### 2.3.2 Multi-objective likelihood function

In this work, we chose three types of eddy covariance data for the calibration: $d_1$ (GPP), $d_2$ (RECO) and $d_3$ (ET). We further assume that the corresponding residual errors, $e_1$, $e_2$ and $e_3$, are uncorrelated, leading to the following multi-objective log-likelihood function

$$\ell(z|d_{1,2,3}) = \ell(z|d_1) + \ell(z|d_2) + \ell(z|d_3).$$

The weighting between the three components of $\ell(z|d_{1,2,3})$ is an important issue. The constant ($\sigma$) and non-constant ($\sigma_i$) standard deviations in equations (3) and (4), respectively, basically weight the respective influences of $e_1$, $e_2$ and $e_3$ on the log-likelihood defined by Eq. (5). Distinct homoscedastic or heteroscedastic residual error models must be specified for $e_1$, $e_2$ and $e_3$. This was done for both the homoscedastic and heteroscedastic cases either by specifying the residual error standard deviations beforehand, or by jointly inferring these standard deviations along with the model parameters.
2.3.3 Homoscedastic and heteroscedastic error models

Based on prior knowledge of the measurement errors, the homoscedasticity assumption simply reduces to assigning values to $\sigma_1$, $\sigma_2$ and $\sigma_3$ in Eqs. (3) and (5). These values were fixed to 3 gC m$^{-2}$ day$^{-1}$ for the GPP measurements, 1.5 gC m$^{-2}$ day$^{-1}$ for the RECO measurements and 1 mm day$^{-1}$ for the ET measurements. As stated earlier, measurement errors associated with eddy covariance fluxes are however typically found to be heteroscedastic, with a variance that is assumed to be linearly related to the magnitude of the measured data $\sigma_{d,i} = \frac{1}{2} \sigma_{0,d} \left( \frac{d_i}{d} + 1 \right)$, \hspace{1cm} (6)

where the variable $d$ denotes either GPP, RECO or ET measurements, $i = 1, \cdots, N$ are measurement times, and $\sigma_{0,d}$ is equivalent to $\sigma_1$, $\sigma_2$, or $\sigma_3$ in the homoscedastic case. We refer to the inversions based on these homoscedastic and heteroscedastic error models as HO1 and HE1, respectively. It is worth noting that by fixing the standard deviations to known measurement errors, one implicitly assumes that the model is able to describe the observed system up to the observation errors. This might not be realistic in environmental modelling where models are always fairly simplified descriptions of a much more complex reality.

2.3.4 Joint inference of the homoscedastic and heteroscedastic error model parameters

Still under the Gaussianity assumption, a more advanced treatment of the residual error models considers simultaneous inference of the standard deviations with the model parameters, i.e., considering the standard deviation of the residual errors as unknowns. Doing so assumes that residual errors are expected to be a mixture of both model (equations and inputs) and observational errors. For the homoscedastic case, this simply consists of jointly sampling $\sigma_1$, $\sigma_2$ and $\sigma_3$ along with the model parameters, $z$.

The heteroscedastic error model then becomes

$\sigma_{d,i} = a d_i + b,$ \hspace{1cm} (7)

where the $a$ and $b$ coefficients are to be jointly inferred with $z$ from the measurement data. Using Eq. (7) thus leads to the addition of 6 variables to the sampling problem: $a_1$, $a_2$, $a_3$, $b_1$, $b_2$ and $b_3$. We refer to the joint inversions of these homoscedastic and heteroscedastic error models as HO2 and HE2, respectively. In these inversions, a total predictive uncertainty around the model output can be computed by adding to the modelled data a random noise drawn from a normal distribution with mean zero and standard deviation $\sigma$ sampled from its posterior distribution (HO2) or computed by Eq. (7) (HE2).

Simultaneous inference of model parameters with homoscedastic or heteroscedastic error model parameters requires the definition of their prior probability distributions. Based on the available prior information, uniform (flat) priors are used for the 10 model parameters contained in $z$ (see Table 2).
We follow two guidelines for specifying the prior densities of the error model parameters. First, we would like to obtain posterior standard deviations as small as possible within the range permitted by the model and measurement data errors in order to get the lowest possible data misfits. Second, the magnitudes of the different prior distributions should reflect the desired weights of the different data types within the multi-objective inference. These weights translate the modeller’s relative preferences among the three modelling objectives in Eq. (5). We therefore use normal distributions with mean zero truncated at zero to avoid negative values. The prescribed weights then correspond to the different standard deviations of these normal distributions

\[ p(X) = \frac{1}{\sigma_X B} \phi \left( \frac{X - \mu_X}{\sigma_X} \right) \propto \phi \left( \frac{X - \mu_X}{\sigma_X} \right), \tag{8} \]

where the \( X \) variable is either \( \sigma_j, a_j \) or \( b_j \) for \( j = 1, 2, 3 \), the value of \( \sigma_X \) expresses the modeller’s preference for objective \( j \) compared to the other objectives (the smaller \( \sigma_X \), the larger the relative weight of objective \( j \)), \( \phi(\cdot) \) signifies the probability density function of the standard normal distribution, \( \mu_X \) is set to zero for maximizing the prior density of \( X \) towards small values, and the constant \( B \) depends on the lower, \( v \), and upper, \( w \), limits of the truncation interval

\[ B = \Phi \left( \frac{w - \mu_X}{\sigma_X} \right) - \Phi \left( \frac{v - \mu_X}{\sigma_X} \right), \tag{9} \]

in which \( \Phi(\cdot) \) denotes the cumulative distribution function of the standard normal distribution.

This treatment of multi-objective Bayesian inference is in line with the work of Reichert and Schuwirth (2012), who further considered different statistical models for model and observation errors. Overall, this resulted in four different ways of treating the eddy covariance data uncertainties: fixed homoscedastic (HO1) and heteroscedastic (HE1) error models, and jointly inferred homoscedastic (HO2) and heteroscedastic (HE2) error models. Using the HO1 and HE1 models led to a total of 10 inferred parameters, whereas using the HO2 and HE2 models resulted into a total of 13 and 16 inferred parameters, respectively. Table 2 lists the marginal prior distributions used for all sampled parameters. The upper and lower bounds of the prior parameter distributions were set using boundary values that correspond at least to the lower and upper physically-possible bounds of the parameters, or to narrower bounds using expert-knowledge.

### 2.3.5 Markov chain Monte Carlo sampling

The goal of the inference is to estimate the posterior distribution \( p(z|d) \) where the 10, 13 or 16-dimensional \( z \) vector contains all sampled parameters, and \( d \) signifies the conditioning data: \( d = \{d_{1,2,3}\} \) herein. As an exact analytical solution of \( p(z|d) \) is not available, we resort to Markov chain Monte Carlo (MCMC) simulation to generate samples from this distribution. The basis of this technique is a Markov chain that generates a random walk through the search space and iteratively finds parameter sets with stable frequencies stemming from the posterior pdf of the model parameters (see, e.g., Roberts 2004 for a comprehensive overview of MCMC simulation).
Table 2. Default values and prior distributions of the 10 model parameters, and prior distributions of the statistical parameters of the homoscedastic and heteroscedastic error models. The label U means an uniform distribution, TG signifies a zero-mean Gaussian distribution truncated at zero to avoid negative values, and SD denotes the prescribed standard deviation of a TG distribution.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Units</th>
<th>Default value</th>
<th>Prior type</th>
<th>Range</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>mol m$^{-2}$ s$^{-1}$</td>
<td>0.01</td>
<td>U</td>
<td>[0.005 − 0.03]</td>
<td>N/A</td>
</tr>
<tr>
<td>g0</td>
<td>mol m$^{-2}$ s$^{-1}$</td>
<td>0.025</td>
<td>U</td>
<td>[0.01 − 0.08]</td>
<td>N/A</td>
</tr>
<tr>
<td>(\tau)</td>
<td>year</td>
<td>6.67</td>
<td>U</td>
<td>[0.5 − 2]</td>
<td>N/A</td>
</tr>
<tr>
<td>(\tau_s)</td>
<td>year</td>
<td>0.0833</td>
<td>U</td>
<td>[0.01 − 0.5]</td>
<td>N/A</td>
</tr>
<tr>
<td>C/N1</td>
<td></td>
<td>16</td>
<td>U</td>
<td>[5 − 40]</td>
<td>N/A</td>
</tr>
<tr>
<td>C/N2</td>
<td></td>
<td>32</td>
<td>U</td>
<td>[10 − 80]</td>
<td>N/A</td>
</tr>
<tr>
<td>(\gamma_1)</td>
<td></td>
<td>20</td>
<td>U</td>
<td>[5 − 40]</td>
<td>N/A</td>
</tr>
<tr>
<td>(\gamma_2)</td>
<td></td>
<td>10</td>
<td>U</td>
<td>[5 − 40]</td>
<td>N/A</td>
</tr>
<tr>
<td>(\gamma_3)</td>
<td></td>
<td>0.2</td>
<td>U</td>
<td>[0 − 1]</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Homoscedastic error model parameters (for HO2 inversions only)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Units</th>
<th>Default value</th>
<th>Prior type</th>
<th>Range</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sigma_{GPP})</td>
<td>gC m$^{-2}$ day$^{-1}$</td>
<td>N/A</td>
<td>TG</td>
<td>[0, 54]</td>
<td>9</td>
</tr>
<tr>
<td>(\sigma_{RECO})</td>
<td>gC m$^{-2}$ day$^{-1}$</td>
<td>N/A</td>
<td>TG</td>
<td>[0, 27]</td>
<td>4.5</td>
</tr>
<tr>
<td>(\sigma_{ET})</td>
<td>mm day$^{-1}$</td>
<td>N/A</td>
<td>TG</td>
<td>[0, 18]</td>
<td>3</td>
</tr>
</tbody>
</table>

Heteroscedastic error model parameters (for HE2 inversions only)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Units</th>
<th>Default value</th>
<th>Prior type</th>
<th>Range</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_{GPP})</td>
<td>gC m$^{-2}$ day$^{-1}$</td>
<td>N/A</td>
<td>TG</td>
<td>[0, 27∗ (Y_{GPP})]</td>
<td>4.5∗ (Y_{GPP})</td>
</tr>
<tr>
<td>(a_{RECO})</td>
<td>gC m$^{-2}$ day$^{-1}$</td>
<td>N/A</td>
<td>TG</td>
<td>[0, 13.5∗ (Y_{RECO})]</td>
<td>2.25∗ (Y_{RECO})</td>
</tr>
<tr>
<td>(a_{ET})</td>
<td>mm day$^{-1}$</td>
<td>N/A</td>
<td>TG</td>
<td>[0, 9∗ (Y_{ET})]</td>
<td>1.5∗ (Y_{ET})</td>
</tr>
<tr>
<td>(b_{GPP})</td>
<td>gC m$^{-2}$ day$^{-1}$</td>
<td>N/A</td>
<td>TG</td>
<td>[0, 27]</td>
<td>4.5</td>
</tr>
<tr>
<td>(b_{RECO})</td>
<td>gC m$^{-2}$ day$^{-1}$</td>
<td>N/A</td>
<td>TG</td>
<td>[0, 13.5]</td>
<td>2.25</td>
</tr>
<tr>
<td>(b_{ET})</td>
<td>mm day$^{-1}$</td>
<td>N/A</td>
<td>TG</td>
<td>[0, 9]</td>
<td>1.5</td>
</tr>
</tbody>
</table>

*Not applicable

The MCMC sampling efficiency strongly depends on the assumed proposal distribution used to generate transitions in the Markov chain. In this work, the state-of-the-art DREAM\((ZS)\) [ter Braak and Vrugt, 2008] [Vrugt et al., 2009] [Laloy and Vrugt, 2012] algorithm is used to generate posterior samples. A detailed description of this sampling scheme including convergence proof can be found in the cited literature and is thus not reproduced herein.

Convergence of the MCMC sampling to the posterior distribution is monitored by means of the potential scale reduction factor of Gelman and Rubin [1992], \(\hat{R}\). This statistic compares for each parameter of interest the average within-chain variance to the variance of all the chains mixed together. The smaller the difference between these two variances, the closer to 1 the value of the \(\hat{R}\) diagnostic. Values of \(\hat{R}\) smaller than 1.2 are commonly deemed to indicate convergence to a stationary distribution. In this study, posterior distributions of the parameters were drawn from the point where all parameters achieved \(\hat{R} < 1.2\). This is more conservative than conventional practice of stopping the inference when \(\hat{R} < 1.2\) for every parameter. The mean acceptance rate of the proposed samples,
AR (%), is an important sampling property and is thus also reported. An excessively small fraction of accepted candidate points indicates poor mixing of the chains due to a too wide proposal distribution. In contrast, a very large acceptance rate signals a too narrow proposal distribution causing the chains to remain in close vicinity of their current locations. The optimal value for AR depends on the proposal and target distributions, but a range of 10-30% generally indicates good performance of DREAM(ZS).

3 Results

3.1 Parameter estimation

3.1.1 Parameter samplings and convergence of the algorithm

Figure 1. Sampled values of the specific leaf area (SLA) by DREAM(ZS) parametrised with 4 chains, for the Oensingen site and the fixed homoscedastic error model (inversion HO1). The vertical dashed line indicates when convergence has been reached according to the $\hat{R}$ statistic.

The DREAM(ZS) algorithm was run with four parallel chains, initialized by sampling the prior parameter distribution (Table 2). As an example, Fig. 1 shows sampling trajectories of DREAM(ZS) parametrised with four chains, for the SLA parameter and inversion HO1 at the Oensingen site. The $\hat{R}$ convergence statistic becomes < 1.2 for each parameter after about 20,000 forward model runs and the AR over the last 50% model evaluations is about 18%. Overall, convergence was achieved for all MCMC trials after some 15,000-30,000 forward runs with AR values in the range 10%-30%, except for the inversions associated with the Laqueuille site that showed AR values as low as 5%.

3.1.2 Posterior parameter distributions

Figure 2 presents marginal posterior histograms of the 10 model parameters for all experimental sites, considering the inferred homoscedastic error model (inversion HO2). In the remainder of this
Figure 2. Posterior distributions of the CARAIB model parameters sampled by the DREAM\textsubscript{(ZS)} algorithm, inferred homoscedastic error model (HO2 inversions), for all sites. The default values (see Table 1) are depicted with a cross and the most likely values with a star. The X-axes cover the whole prior ranges.
document, results are mainly detailed for this inversion scenario, since it generally led to the lowest data misfit statistics in calibration. For some parameters (e.g., SLA and C/N1), the marginal posterior distributions are narrow compared to the prior parameter range. This indicates a large sensitivity of the model to the considered parameter. In contrast, some other parameters such as $\gamma^2$ are poorly resolved, demonstrating a relative insensitivity. Asymmetric edge-hitting distributions are also observed such as for C/N1 and C/N2 in Monte-Bondone. In a Bayesian inversion of eddy covariance data obtained from a forest site, Braswell et al. (2005) found that 7 out of 26 marginal parameter distributions were edge-hitting. Extending the prior parameter ranges would lead herein to unphysical or unplausible parameter values. Edge-hitting distributions reveal model inadequacies and/or large systematic measurements errors. For some parameters, posterior distributions were rather distinct from the default values that were used in previous studies (Table 2), such as high g1 values. Values of the characteristic mortality time $\tau$ also generally increased compared to the default value.

Table 3 shows for the four experimental sites the most likely parameter values, which resulted in the highest values of the log-likelihood function. Some of the parameters present contrasting values between inversion scenarios and/or experimental sites, which may be related to the different ecological characteristics of the sites as discussed in section 4.3. Depending on the width of the posterior distributions, the most likely parameter values are well resolved or largely uncertain. As a result, comparison between the experimental sites must account for the posterior distributions of the parameters.

3.2 Measured and modelled carbon and water fluxes with calibration data

3.2.1 Measured and modelled data in Monte-Bondone

As the parameters sampling resulted in posterior distributions of the parameters instead of single values, ensembles of posterior modelled signals can be represented. In Fig. 3 measured and modelled eddy covariance data are depicted for the experimental site of Monte-Bondone, for inversions with the inferred homoscedastic error model (inversion HO2). The posterior ranges of the modelled signals are represented by the dark grey shaded areas for the prediction uncertainty due to parameter uncertainties and by the light grey shaded areas for the total predictive uncertainty (at 95% confidence interval). This total prediction uncertainty is computed using the standard deviation of the residual errors $\sigma$ as sampled by the inversions and, therefore, cannot be computed for the NEE. The site of Monte-Bondone was chosen here since there is one single cut a year (indicated by the vertical arrows) that is clearly identifiable, which facilitates the interpretation of the fluxes. The dates of cutting corresponded to a sudden drop in the GPP in the middle of the year, that was followed by a gradual increase. They were also observed in the NEE graphs with a sudden increase in the NEE.

There were overall good agreements between measured and modelled signals. It is worth noting that the posterior ranges of modelled data were not constant over time and were not related to the
Table 3. Most likely CARAIB model parameters values for all inversion scenarios

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Grilleryburg</th>
<th>Oensingen</th>
<th>Monte-Bondone</th>
<th>Laqueuille</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed homoscedastic error model inversions (HO1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g1</td>
<td>16.8</td>
<td>7.3</td>
<td>18.8</td>
<td>18.6</td>
</tr>
<tr>
<td>g0 [mol m$^{-2}$ s$^{-1}$]</td>
<td>0.0265</td>
<td>0.00507</td>
<td>0.00637</td>
<td>0.0248</td>
</tr>
<tr>
<td>SLA [m$^2$gC$^{-1}$]</td>
<td>0.0126</td>
<td>0.0234</td>
<td>0.0155</td>
<td>0.0197</td>
</tr>
<tr>
<td>$\tau$ [year]</td>
<td>1.99</td>
<td>1.27</td>
<td>1.98</td>
<td>1.49</td>
</tr>
<tr>
<td>$\tau_s$ [year]</td>
<td>0.0861</td>
<td>0.0526</td>
<td>0.0212</td>
<td>0.023</td>
</tr>
<tr>
<td>C/N1</td>
<td>5</td>
<td>6.69</td>
<td>5.02</td>
<td>5.43</td>
</tr>
<tr>
<td>C/N2</td>
<td>78.6</td>
<td>19.9</td>
<td>10.6</td>
<td>11</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>5.07</td>
<td>39.1</td>
<td>38.2</td>
<td>26.1</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>5.1</td>
<td>39.9</td>
<td>38.8</td>
<td>36.9</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>0.73</td>
<td>0.507</td>
<td>0.421</td>
<td>1.49e-05</td>
</tr>
<tr>
<td>Fixed heteroscedastic error model inversions (HE1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g1</td>
<td>3.45</td>
<td>8</td>
<td>19.8</td>
<td>20</td>
</tr>
<tr>
<td>g0 [mol m$^{-2}$ s$^{-1}$]</td>
<td>0.027</td>
<td>0.00544</td>
<td>0.0297</td>
<td>0.0299</td>
</tr>
<tr>
<td>SLA [m$^2$gC$^{-1}$]</td>
<td>0.0161</td>
<td>0.0151</td>
<td>0.0142</td>
<td>0.0191</td>
</tr>
<tr>
<td>$\tau$ [year]</td>
<td>1.96</td>
<td>1.7</td>
<td>1.96</td>
<td>0.746</td>
</tr>
<tr>
<td>$\tau_s$ [year]</td>
<td>0.0202</td>
<td>0.0687</td>
<td>0.0153</td>
<td>0.0234</td>
</tr>
<tr>
<td>C/N1</td>
<td>5.11</td>
<td>5.1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>C/N2</td>
<td>77.9</td>
<td>20.3</td>
<td>10.2</td>
<td>10</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>8.09</td>
<td>39.5</td>
<td>31.4</td>
<td>38.8</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>5.96</td>
<td>37.4</td>
<td>30.9</td>
<td>24.8</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>0.358</td>
<td>0.806</td>
<td>0.981</td>
<td>0.688</td>
</tr>
<tr>
<td>Inferred homoscedastic error model inversions (HO2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g1</td>
<td>15.6</td>
<td>7.46</td>
<td>16.8</td>
<td>14.5</td>
</tr>
<tr>
<td>g0 [mol m$^{-2}$ s$^{-1}$]</td>
<td>0.00945</td>
<td>0.00549</td>
<td>0.0258</td>
<td>0.0104</td>
</tr>
<tr>
<td>SLA [m$^2$gC$^{-1}$]</td>
<td>0.0133</td>
<td>0.0193</td>
<td>0.0142</td>
<td>0.0483</td>
</tr>
<tr>
<td>$\tau$ [year]</td>
<td>1.98</td>
<td>1.65</td>
<td>1.99</td>
<td>0.65</td>
</tr>
<tr>
<td>$\tau_s$ [year]</td>
<td>0.0682</td>
<td>0.0583</td>
<td>0.0735</td>
<td>0.0102</td>
</tr>
<tr>
<td>C/N1</td>
<td>5.57</td>
<td>5.43</td>
<td>5</td>
<td>15.6</td>
</tr>
<tr>
<td>C/N2</td>
<td>77</td>
<td>20.2</td>
<td>10</td>
<td>52.7</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>6.25</td>
<td>37.2</td>
<td>20.8</td>
<td>39.6</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>5.26</td>
<td>35.5</td>
<td>27.3</td>
<td>5.58</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>0.257</td>
<td>0.471</td>
<td>0.361</td>
<td>0.000272</td>
</tr>
<tr>
<td>Inferred heteroscedastic error model inversions (HE2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g1</td>
<td>11.3</td>
<td>9.4</td>
<td>19.8</td>
<td>12.7</td>
</tr>
<tr>
<td>g0 [mol m$^{-2}$ s$^{-1}$]</td>
<td>0.0276</td>
<td>0.00635</td>
<td>0.0298</td>
<td>0.0234</td>
</tr>
<tr>
<td>SLA [m$^2$gC$^{-1}$]</td>
<td>0.018</td>
<td>0.0158</td>
<td>0.0142</td>
<td>0.0797</td>
</tr>
<tr>
<td>$\tau$ [year]</td>
<td>1.69</td>
<td>1.8</td>
<td>1.27</td>
<td>0.822</td>
</tr>
<tr>
<td>$\tau_s$ [year]</td>
<td>0.01</td>
<td>0.0892</td>
<td>0.0141</td>
<td>0.0104</td>
</tr>
<tr>
<td>C/N1</td>
<td>6.67</td>
<td>5.4</td>
<td>5.1</td>
<td>20.6</td>
</tr>
<tr>
<td>C/N2</td>
<td>22.9</td>
<td>15.5</td>
<td>14.9</td>
<td>10.1</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>7.83</td>
<td>21.7</td>
<td>20.6</td>
<td>38.5</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>6.14</td>
<td>21.9</td>
<td>19.1</td>
<td>9.79</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>0.503</td>
<td>0.896</td>
<td>0.145</td>
<td>0.505</td>
</tr>
</tbody>
</table>
Figure 3. Measured and modelled GPP [gC m\(^{-2}\) day\(^{-1}\)] (a), RECO [gC m\(^{-2}\) day\(^{-1}\)] (b), ET [mm day\(^{-1}\)] (c) and NEE [gC m\(^{-2}\) day\(^{-1}\)] (d) at the Monte-Bondone site for the inferred homoscedastic error model (inversion HO2). The ranges of the prediction uncertainty due to parameter uncertainties and the 95% total predictive uncertainty (only for GPP, RECO and ET) are depicted by the dark and light grey shaded areas, respectively. Vertical arrows indicate the dates of the grass cutting.
magnitude of the signals. The ranges due to parameter uncertainties were relatively small and did not encompass the measured data. Overall, it could be observed that measured eddy covariance data have a stronger kinetic than the modelled signals, meaning that the CARAIB model cannot follow the fast fluctuations of the GPP (and other signals) over time. In particular, the model could not well simulate the highest peaks of GPP.

3.2.2 Measured and modelled data across sites

Considering the other three experimental sites (Fig. 4), there were similar agreements between measured and modelled signals, although the sites displayed different behaviour in terms of GPP as their management is varying: there are several cuts per year in Grillenburg and Oensingen, while Laqueuille is a grazed meadow. In general, the peaks of GPP cannot be well simulated by the model. The modelled GPP seemed averaged out as compared to the measured signals, as observed before in Monte-Bondone (Fig. 3a).

Table 4. Comparison between measured and modelled signals using most likely parameter values. \(ml\) is the maximum value of the log-likelihood function.

<table>
<thead>
<tr>
<th></th>
<th>Grillenburg</th>
<th>Oensingen</th>
<th>Monte-Bondone</th>
<th>Laqueuille</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ml)</td>
<td>-5560</td>
<td>-7402</td>
<td>-5248</td>
<td>-8284</td>
</tr>
<tr>
<td>GPP ([\text{gC m}^{-2} \text{day}^{-1}])</td>
<td>1.797 0.726 0.791</td>
<td>2.231 0.600 0.757</td>
<td>1.742 0.755 0.831</td>
<td>2.151 0.521 0.751</td>
</tr>
<tr>
<td>RECO ([\text{gC m}^{-2} \text{day}^{-1}])</td>
<td>1.498 0.502 0.695</td>
<td>1.269 0.772 0.803</td>
<td>1.036 0.832 0.878</td>
<td>1.529 0.688 0.743</td>
</tr>
<tr>
<td>ET ([\text{mm day}^{-1}])</td>
<td>0.623 0.309 0.565</td>
<td>0.670 0.612 0.758</td>
<td>0.500 0.784 0.849</td>
<td>1.128 0.144 0.474</td>
</tr>
<tr>
<td>NEE ([\text{gC m}^{-2} \text{day}^{-1}])</td>
<td>1.774 -0.185 0.335</td>
<td>2.044 -0.115 0.449</td>
<td>1.424 -0.018 0.463</td>
<td>2.153 -0.382 0.219</td>
</tr>
</tbody>
</table>

All the graphical comparisons between measured and modelled signals could not be shown, but are summarized in Table 4 for the homoscedastic and heteroscedastic cases, and with fixed and inferred error model, using the root mean square error (RMSE), the \(R^2\) and the Nash and Sutcliffe...
Figure 4. Measured and modelled GPP [g C m$^{-2}$ day$^{-1}$] for the Grillenburg (a), Oensingen (b) and Laqueuille (c) experimental sites. See Fig. 3(a) for Monte-Bondone. The ranges of the prediction uncertainty due to parameter uncertainties and the 95% total predictive uncertainty are depicted by the dark and light grey shaded areas, respectively. Vertical arrows indicate the dates of the grass cutting (Grillenburg and Oensingen) and horizontal arrows the periods of grazing (Laqueuille).
model efficiency criterion (E) between measured and modelled signals. The latter criterion takes values from $-\infty$ to 1. A value of 1 means a perfect match between measurements and model simulations, a value of 0 indicates that the mean of the observed data is as accurate as the modelled values, and an efficiency less than 0 occurs when the mean of the observed data reproduces the observations better than the modelled values. The maximum log-likelihood value $m_l$ that was obtained by the algorithm is also indicated. Note that performance criteria were also computed for the NEE, although these data were not used in the model inversions. Overall, the best agreements were found for the Monte-Bondone site, and the worst for the Laqueuille site. The lowest model efficiencies $E$ were found for the NEE, which is not surprising since these data were not accounted for in the model inversions. While the $m_l$ values were generally the highest for the heteroscedastic inversions HE2, RMSE appeared larger for these inversions.

### 3.2.3 Homoscedastic and heteroscedastic eddy covariance residual errors

Considering homoscedastic or heteroscedastic residual eddy covariance residual errors resulted in different sampling of posterior distributions of parameters, and therefore, different posterior modelled signals of the model. As an example, Fig. 5 shows the measured and modelled GPP with their posterior ranges for the site of Monte-Bondone in 2004, for both homoscedastic (a,c) and heteroscedastic (b,d) cases. For the HO2 and HE2 inversions, the 95% total predictive uncertainty is depicted using the light grey shaded areas. The measurement uncertainty is depicted only for fixed eddy covariance residual errors inversions (a,b) for clarity. The measurement uncertainty is thus constant for the homoscedastic case (namely, $\pm 3 \text{ gC m}^{-2} \text{ day}^{-1}$ for HO1) while it varies linearly according to the GPP for the heteroscedastic case (HE1). This two options led to different behaviours of the modelled GPP using the posterior distributions, which better approached the high values of the measured data (in summer) in the homoscedastic cases and better fit the low values (in winter) in the heteroscedastic cases. Overall, in calibration, modelled signals with parameters values from the homoscedastic inversions were in a better agreement with the measured data than with the parameters from the heteroscedastic inversions. The same observation was also made for the other sites (not shown), as it could also be observed in Table 4. However, the total predictive uncertainty range derived from the HE2 inversions was more consistent, as, e. g., it avoids unrealistic negative values of GPP. The standardised residuals, that were computed as the difference between measured and modelled data divided by the standard deviation of the residual error, are depicted in Fig. 5 at the right of the GPP graphs. Heteroscedasticity of the GPP residual errors was fairly reduced but not fully removed by using the HE1 and HE2 heteroscedastic residual error models. Indeed, the standardised residuals still showed some small but complex heteroscedastic patterns. Partial autocorrelation of the residuals of the GPP were also depicted and independence between the days of simulation was reached after a few days.
Figure 5. Measured and modelled GPP [gC m$^{-2}$ day$^{-1}$] at the Monte-Bondone site in 2004 for the fixed homoscedastic HO1 (a) and heteroscedastic HE1 (b), inferred homoscedastic HO2 (c) and heteroscedastic HE2 (d) inversions. The measured GPP is depicted with a constant (a) and variable (b) uncertainty range. For the HO2 and HE2 inversions, the 95% confidence interval total predictive uncertainty is depicted using the light grey shaded areas. Standardised residuals and partial autocorrelation of residuals of GPP over the full simulation period are depicted at the right of each graph.
3.2.4 Sampling of the standard deviation of the residual errors

| Table 5. Most likely standard deviation of the residual errors (HO2) and parameters of Eq. (7) (HE2). |
|-------------------------------------------------|---------------------------------|----------------|----------------|----------------|
| Inferred homoscedastic inversions (HO2)         | Grillenburg | Oensingen | Monte-Bondone | Laqueuille   |
| \( \sigma_{GPP} \)                             | 1.81        | 2.29      | 1.79          | 2.22         |
| \( \sigma_{RECO} \)                            | 1.63        | 1.33      | 1.09          | 1.62         |
| \( \sigma_{ET} \)                              | 0.632       | 0.682     | 0.519         | 1.31         |
| Inferred heteroscedastic inversions (HE2)       |             |           |               |              |
| \( a_{GPP} \)                                  | 0.211       | 0.65      | 0.336         | 1.09         |
| \( a_{RECO} \)                                 | 0.12        | 0.334     | 0.162         | 0.514        |
| \( a_{ET} \)                                   | 0.246       | 0.255     | 0.316         | 0.818        |
| \( b_{GPP} \)                                  | 0.406       | 0.297     | 0.423         | 0.239        |
| \( b_{RECO} \)                                 | 0.411       | 0.206     | 0.283         | 0.233        |
| \( b_{ET} \)                                   | 0.273       | 0.255     | 0.12          | 0.175        |

Inversions with the sampling of the standard deviations of the residual errors resulted in posterior distributions of the standard deviation of the residual errors (HO2) and parameters of Eq. (7) (HE2). Most likely values of these distributions (Table 5) were depending on the experimental sites, being larger for Laqueuille and Oensingen, which can be related to the poorer agreements between measured and modelled data in these sites. Although the sampled standard deviations of the residual errors were lower than in the fixed inversions, there were no large differences between the inversions with fixed model errors (HO1 & HE1) and inversions with inferred model errors (HO2 & HE2) in terms of agreement between measured and modelled signals (see Fig. 5 and Table 4) or in the posterior distributions of parameters (Table 3).

3.3 Model validation

Parameter values from the posterior distributions were tested for validation using eddy covariance data over different periods (validation datasets, see Table 1). Figure 6 shows measured and modelled GPP values over the periods of calibration and validation in Monte-Bondone. Not surprisingly, worse agreements between measured and modelled data are observed as compared to the calibration period. However, it is observed that the modelled GPP in validation in the HE2 inversions follows better the measured signal than in the HO2 inversions. Strikingly, in all the sites, the posterior parameter distributions derived from using the HE1 and HE2 heteroscedastic models are found to induce a better model performance in validation compared to the posterior distributions associated with the use of the homoscedastic models (Table 6). The difference between calibration and validation appeared thus smaller when using most likely parameter values from heteroscedastic inversions as compared to homoscedastic inversions. Among the different grassland sites, a similar performance pattern as for the calibration experiment is observed. Indeed, the Laqueuille site shows for each type of mea-
asurement data the worst performance statistics whereas the Monte-Bondone site overall presents the best fits to the data (Table 6).

Table 6. Validation of the calibrated model using most likely parameter values from the inversions

<table>
<thead>
<tr>
<th></th>
<th>Grillenburg</th>
<th>Oensingen</th>
<th>Monte-Bondone</th>
<th>Laqueuille</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed homoscedastic error model inversions (HO1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPP [gC m(^{-2}) day(^{-1})]</td>
<td>2.647</td>
<td>0.581</td>
<td>0.694</td>
<td>3.000</td>
</tr>
<tr>
<td>RECO [gC m(^{-2}) day(^{-1})]</td>
<td>1.284</td>
<td>0.768</td>
<td>0.799</td>
<td>1.560</td>
</tr>
<tr>
<td>ET [mm day(^{-1})]</td>
<td>0.642</td>
<td>0.520</td>
<td>0.602</td>
<td>0.732</td>
</tr>
<tr>
<td>NEE [gC m(^{-2}) day(^{-1})]</td>
<td>1.800</td>
<td>0.197</td>
<td>0.453</td>
<td>2.332</td>
</tr>
<tr>
<td><strong>Fixed heteroscedastic error model inversions (HE1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPP [gC m(^{-2}) day(^{-1})]</td>
<td>2.030</td>
<td>0.716</td>
<td>0.748</td>
<td>2.803</td>
</tr>
<tr>
<td>RECO [gC m(^{-2}) day(^{-1})]</td>
<td>1.221</td>
<td>0.790</td>
<td>0.871</td>
<td>1.555</td>
</tr>
<tr>
<td>ET [mm day(^{-1})]</td>
<td>0.610</td>
<td>0.567</td>
<td>0.613</td>
<td>0.678</td>
</tr>
<tr>
<td>NEE [gC m(^{-2}) day(^{-1})]</td>
<td>1.972</td>
<td>0.117</td>
<td>0.281</td>
<td>2.206</td>
</tr>
<tr>
<td><strong>Inferred homoscedastic error model inversions (HO2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPP [gC m(^{-2}) day(^{-1})]</td>
<td>2.651</td>
<td>0.516</td>
<td>0.663</td>
<td>3.008</td>
</tr>
<tr>
<td>RECO [gC m(^{-2}) day(^{-1})]</td>
<td>1.335</td>
<td>0.749</td>
<td>0.777</td>
<td>1.614</td>
</tr>
<tr>
<td>ET [mm day(^{-1})]</td>
<td>0.639</td>
<td>0.525</td>
<td>0.603</td>
<td>0.710</td>
</tr>
<tr>
<td>NEE [gC m(^{-2}) day(^{-1})]</td>
<td>2.010</td>
<td>0.082</td>
<td>0.451</td>
<td>2.337</td>
</tr>
<tr>
<td><strong>Inferred heteroscedastic error model inversions (HE2)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPP [gC m(^{-2}) day(^{-1})]</td>
<td>2.361</td>
<td>0.617</td>
<td>0.699</td>
<td>2.825</td>
</tr>
<tr>
<td>RECO [gC m(^{-2}) day(^{-1})]</td>
<td>1.220</td>
<td>0.790</td>
<td>0.815</td>
<td>1.549</td>
</tr>
<tr>
<td>ET [mm day(^{-1})]</td>
<td>0.617</td>
<td>0.557</td>
<td>0.614</td>
<td>0.683</td>
</tr>
<tr>
<td>NEE [gC m(^{-2}) day(^{-1})]</td>
<td>1.837</td>
<td>0.233</td>
<td>0.364</td>
<td>2.227</td>
</tr>
</tbody>
</table>

4 Discussions

4.1 Measured and modelled signals

Bayesian inversions over the four grassland sites resulted in posterior distributions of parameters and posterior ranges of modelled signals (GPP, RECO, ET and NEE). Considering the inversion scenario HO2, there were in general good agreements between measured and modelled signals, with RMSE ranging from 1.73 to 2.19 gC m\(^{-2}\)day\(^{-1}\) and \(R^2\) between 0.74 and 0.84 in terms of GPP. Using a dedicated model for soil organic carbon dynamics, De Bruijn et al. (2012) found a \(R^2\) of 0.68 for the modelling of the NEE at the Oensingen site over the same years. Comparing three large-scale lands surface models in simulating carbon fluxes over different ecosystems, Balzarolo et al. (2014) noticed that grassland and crop sites were more difficult to model compared to forest sites. Using data from 13 grassland sites over Europe including Laqueuille and Grillenburg, they found average RMSE between measured and modelled GPP ranging from 2.45 to 3.57 gC m\(^{-2}\)day\(^{-1}\) and \(R^2\) from 0.37 to 0.56. These larger discrepancies compared to our study are mainly to be related to the fact that the large-scale models were used without site-calibrations. Modelling of carbon fluxes was also performed at the Oensingen site over the same years in Calanca et al. (2007) using a
Figure 6. Measured and modelled GPP [gC m\(^{-2}\) day\(^{-1}\)] at the Monte-Bondone site in calibration (2003-2005) and validation (2006-2007) for the inferred homoscedastic HO2 (a) and heteroscedastic HE2 (b) inversions. The 95% confidence interval total predictive uncertainty is depicted using the light grey shaded areas.

It could be observed that measured eddy covariance data have a stronger kinetic than the modelled signals, that is, modelled signals could not follow the fast fluctuations of the measured signals and, in particular, simulate high GPP values. This could be related to the different time resolutions between the model and data. The CARAIB model is based on daily-averaged meteorological data. However, photosynthesis and respiration processes are computed at a two-hour time step before being aggregated to a daily resolution and the model assumes a symmetry with respect to solar noon time (Otto et al., 2002) to save computation resource. Moreover, in the CARAIB model, solar fluxes are cal-
culated assuming a constant cloudiness over the day and temperature is varying using a sinusoidal function between the minimal and maximal temperature, that were fixed at midnight and noon, respectively. These shortcomings were necessary for saving computation resources and in case of data scarcity for global vegetation modelling. Eddy covariance data, however, are typically acquired at a time frequency of 5 or 10 Hz (Aubinet et al., 2012) and can thus capture high-frequency fluxes. Even though eddy covariance data were aggregated over time to a daily time resolution, the high-frequency acquisition rate ensures that effects of abrupt meteorological events are recorded. Increasing the time resolution of the CARAIB model would help to better simulate ecophysiological processes at a high frequency. Alternatively, a simple workaround to deal with the different time dynamics would be to apply a filter based on a moving window of some days in order to smooth measured (and modelled) eddy covariance data before computing the statistical indicators, as done in Calanca et al. (2007).

Another modelling limitation is that model parameters are assumed as constant along the season, although plants traits are known to evolve throughout the season and plants acclimate to specific climate conditions. As a result, the effect of similar climatic conditions does not necessary result in similar eddy covariance measurements.

In general, there were poorer agreements between measured and modelled signals (GPP, RECO, ET and NEE) in Laqueuille compared to the other experimental sites. These poorer agreements can be probably related to the grazing instead of the cutting that occurs in Laqueuille. Grazing was more difficult to simulate because of the expert-knowledge conversion between the given cattle charge and the biomass removal. As a result, grass cutting is better constrained in the model compared to grazing, as it was already shown in the Laqueuille experimental site by Calanca et al. (2007) but using the grassland model PaSim.

All the same, besides the average statistical indicators between measured and modelled signals, the performance of the calibration might be also evaluated against specific scientific or operational objectives. For instance, accurate modelling of the grass cutting or computation of annual budgets of carbon in the grassland (e.g., Soussana et al., 2007) might show different performances, depending on the time scale on which the processes are analysed.

### 4.2 Eddy covariance residual errors

#### 4.2.1 Homoscedastic and heteroscedastic eddy covariance residual errors

Bayesian inversions were conducted considering homoscedasticity and heteroscedasticity in the eddy covariance residual errors. Figure [5] showed that accounting for heteroscedasticity in eddy covariance residual errors permitted to better simulate low-magnitude signals (winter), but at the same time, it penalized the modelling of high-magnitude signals (summer). Actually, it is worth remarking that inversions considering heteroscedastic measurement errors do not attempt to result in smaller misfits between measured and modelled data since larger errors are considered for high
peaks of the signals. However, in validation, the posterior parameter distributions derived from using the heteroscedastic residual error models outperform their counterparts derived from using the homoscedastic residual error models. This important finding reveals that despite inducing larger RMSE values in calibration, the use of a heteroscedastic residual error model leads to a more robust parameter estimation.

Since eddy covariance data are known to show heteroscedasticity, accounting for a heteroscedastic model of the residuals errors in the inversions is more conceptually sound for ensuring unbiased posterior distributions of parameters. However, we showed that considering a linear heteroscedastic model of the residual errors only partly removed heteroscedasticity in the standardised residuals values (Fig. 5 (b) and (d)). Other kinds of heteroscedastic models (i.e., non-linear) might be tested, but the residual distributions did not show any clear trend for all sites.

It is also worth noting that a substantial fraction of the large residual errors is caused by the tendency of the CARAIB model of underestimating the observed GPP summer peaks. As discussed above, this is related to a slower temporal resolution of the model compared to that of the measured data. To overcome this model inadequacy, further model modifications are necessary to increase the time resolution of the model. Another model improvement would be to simulate varying model parameter values as a function of the time of the year, since plant traits are actually evolving along the seasons. However, this would come at the cost of a large increase in model complexity.

4.2.2 Sampling of the standard deviation of residual errors

Sampling the standard deviation of residual errors, i.e., the inversions HO2 and HE2, did not impact a lot the other parameter samplings and the modelling, as compared to inversions HO1 and HE1, respectively. Some performance criteria were better with the sampling of the residual standard deviations, while other not. As expected, most likely standard deviation of the residuals errors were close to the RMSE obtained in the inversions HO2. The benefit of these values is that they inform about the levels of the uncertainties of the eddy covariance data with respect with the model used to invert the data, e.g., uncertainties of GPP ranged from 1.79 to 2.29 gC m$^{-2}$ day$^{-1}$, of RECO from 1.09 to 1.63 gC m$^{-2}$ day$^{-1}$ and of ET from 0.52 to 1.31 mm day$^{-1}$. They could be used to weight different eddy covariance data in multi-objective inverse modelling.

4.3 Parameters values across sites

Posterior distributions of parameters showed contrasting values that could be linked to the characteristics of the experimental sites. For instance, the specific leaf area (SLA) is known to depend on many factors (Marcelis et al., 1998) such as leaf age, temperature, light intensity, aridity and soil nutrient content. Thick leaves (low SLA) are more adapted to dry ecosystems due to their greater capacity to retain water. Although none of the 4 grassland sites are strictly characterized by a dry climate, it is interesting to note that the posterior parameter distributions for SLA were negatively
correlated with the aridity, inversely expressed by the De Martonne-Gottman index (Fig. 7), that is, SLA decreases with increasing aridity. The largest SLA (thin leaves) were found for Laqueuille, which can be related to the permanent grazing that constantly regenerates young leaves, since young leaves are characterised by high SLA. The large SLA values in Oensigen can be related to more intensive management conditions (fertilisation, more frequent cuts).

Contrarily to SLA, the characteristic mortality time in stress conditions $\tau_s$ appeared to be positively correlated with the site aridity (Fig. 7). Larger $\tau_s$ value means a larger water stress resistance for the plants in Grillenburg and Monte-Bondone.

The values of $g1$ were drastically different between Oensingen and the three other sites (Table 3). In addition, for these three sites, the values appeared much higher compared to the default values ($g1 = 9$) and other values commonly encountered in the literature [Van Wijk et al., 2000; Medlyn et al., 2011]. It is known that $g1$ should increase with humid conditions and temperature (Medlyn et al., 2011), as it is positively related to the marginal water cost of carbon gain. However, the high values of $g1$ here could not be really related to a warmer or wetter climate as compared to Oensingen. A possible explanation could be related to the different dynamics of the model and the measurements, as already explained hereinbefore. As the model cannot simulate the high GPP values.
that are observed in the eddy covariance data, the Bayesian algorithm could have compensated by
sampling high values of \( g_1 \) that increase stomatal conductance.

More broadly, ecophysiological differences between the grassland sites resulted in posterior dis-
tributions of parameters that can be either drastically different or common between the sites (Fig. 2).
If it appears that site-specific parameter values are needed, it means that the model has to be refined
by accounting for ecophysiological dependence of the parameters. If not, generalised parameters
values could be used, meaning that they are invariant of the site on which they were determined or
even independent from the plant species, as recently claimed by Yuan et al. (2014). Determining a
common set of the parameters distributions among the four sites could be done either by (1) merging
the four posterior distributions after independent samplings of the data of each site or (2) merging
together the eddy covariance data of the four sites in one single MCMC sampling, as explored in
Kuppel et al. (2012).

5 Conclusions

Bayesian inversions of the CARAIB dynamic vegetation model were performed using eddy covari-
ance data (GPP, RECO, ET) at four experimental grassland sites. A specific version of the CARAIB
model was developed for this application, with functions related to the grassland management, i.e.,
grass cutting and grazing. Posterior parameter and predictive distributions were compared for dif-
ferent statistical models of the eddy covariance residual errors: (1) assuming homoscedasticity or
heteroscedasticity of the residual errors, and (2) fixing beforehand or jointly inferring the variances
of the residual errors. There were in general good agreements between measured and modelled sig-
nals for the calibration datasets with RMSE of daily gross primary productivity (GPP), ecosystem
respiration (RECO) and evapotranspiration (ET) ranging from 1.73 to 2.19 gC m\(^{-2}\) day\(^{-1}\), 1.04 to
1.56 gC m\(^{-2}\) day\(^{-1}\), and 0.50 to 1.28 mm day\(^{-1}\) respectively. Since the four sites belong to a long-
standing network of eddy covariance data measurements, comparisons with previous studies could
be made.

Although the eddy covariance measurements errors are known to be heteroscedastic, the use of
a homoscedastic error model led to a better model performance in calibration compared to using
a heteroscedastic error model. Nevertheless, a model validation experiment revealed that CARAIB
models calibrated by means of a heteroscedastic error model outperform those calibrated assuming
homoscedastic residual errors. Posterior parameter distributions derived from using a heteroscedastic
model of the residuals are therefore more sound and robust, even though heteroscedasticity could not
be fully removed. Therefore, our results support the use of a heteroscedastic residual error model for
inverting eddy covariance data and inferring posterior parameter distributions.

Systematic model-data discrepancies were also found for the largest observed GPP values. This
can be attributed to the low temporal resolution of the photosynthetic processes in the CARAIB
model, among other model inadequacies. Modelling performance varied among the four sites, with poorer performances at Laqueuille, because of the greater difficulty of modelling grazing compared to grass-cutting. Lastly, site-specific posterior parameter distributions obtained for the four grasslands were compared and discussed with respect to grassland characteristics. Specific leaf area and characteristic mortality time parameters appeared to be related to site aridity.

Acknowledgements. This research was funded by the “Direction Générale Opérationelle de l’Économie, de l’Emploi & de la Recherche” (DGO6), Wallonie, Belgium. This study was made in the framework of the FACCE/MACSUR knowledge hub, a pan-European collaborative project based on the modelling of agriculture systems facing climate change. In particular, this study fits in the intercomparison of grassland models task, led by Gianni Bellocchi, that aims to compare and improve the performance of eight models in grassland growth simulations.
References


