Earth Syst. Dynam., 7, 917–935, 2016
www.earth-syst-dynam.net/7/917/2016/
doi:10.5194/esd-7-917-2016

Earth System Dynamics
Open Access

The impact of structural error on parameter constraint in a climate model

Doug McNeall[1], Jonny Williams[2,4], Ben Booth[1], Richard Betts[1], Peter Challenor[3], Andy Wiltshire[1], and David Sexton[1]

[1]Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK
[2]BRIDGE, School of Geographical Sciences, University of Bristol, Bristol, BS8 1SS, UK
[3]University of Exeter, North Park Road, Exeter, EX4 4QE, UK
[4]Now at NIWA, 301 Evans Bay Parade, Hataitai, Wellington 6021, New Zealand

Correspondence to: Doug McNeall (doug.mcneall@metoffice.gov.uk)

Received: 18 April 2016 – Published in Earth Syst. Dynam. Discuss.: 29 April 2016
Revised: 7 November 2016 – Accepted: 7 November 2016 – Published: 24 November 2016

Abstract. Uncertainty in the simulation of the carbon cycle contributes significantly to uncertainty in the projections of future climate change. We use observations of forest fraction to constrain carbon cycle and land surface input parameters of the global climate model FAMOUS, in the presence of an uncertain structural error.

Using an ensemble of climate model runs to build a computationally cheap statistical proxy (emulator) of the climate model, we use history matching to rule out input parameter settings where the corresponding climate model output is judged sufficiently different from observations, even allowing for uncertainty.

Regions of parameter space where FAMOUS best simulates the Amazon forest fraction are incompatible with the regions where FAMOUS best simulates other forests, indicating a structural error in the model. We use the emulator to simulate the forest fraction at the best set of parameters implied by matching the model to the Amazon, Central African, South East Asian, and North American forests in turn. We can find parameters that lead to a realistic forest fraction in the Amazon, but that using the Amazon alone to tune the simulator would result in a significant overestimate of forest fraction in the other forests. Conversely, using the other forests to tune the simulator leads to a larger underestimate of the Amazon forest fraction.

We use sensitivity analysis to find the parameters which have the most impact on simulator output and perform a history-matching exercise using credible estimates for simulator discrepancy and observational uncertainty terms. We are unable to constrain the parameters individually, but we rule out just under half of joint parameter space as being incompatible with forest observations. We discuss the possible sources of the discrepancy in the simulated Amazon, including missing processes in the land surface component and a bias in the climatology of the Amazon.

Copyright statement

1 Introduction

Earth system processes that are too high resolution or complex to model explicitly are often simplified or *parameterised*, with tuneable coefficients that quantitatively represent some aspect of the process. The coefficients may represent a measurable physical quantity, or they may be a more abstract representation necessary due to the simplification of the modelled process. Uncertainty about the best value of the coefficients means it may not be desirable to choose a single value over all others. This uncertainty can be represented, for example, by using a range of values for each of the coefficients in an ensemble of simulator[1] runs.

Choosing parameterisation coefficients is a major research effort encompassing domain specific, statistical and computational literature. Coefficients are tuneable by comparing the simulator with observations of the system, by direct measurement or from information from theory. There is a long history of using observations to constrain parameterisation coefficients within general circulation models (GCMs), particularly within atmospheric components. Where this is done in a formal probabilistic setting it can provide probability distributions for the parameters of the simulator; this is known as *calibration*. Choosing a single best parameter set is *tuning*. *History matching* rules out parameter settings where simulator output is statistically inconsistent with observations, given uncertainty in those observations, uncertainty in knowledge of the simulator, and a given tolerance of error. A well-calibrated simulator should match the underlying dynamics of a system better and should produce more accurate and (appropriately) tightly constrained predictions.

1.1 Simulator discrepancy

Simulator discrepancy is the systematic difference between a climate model, or simulator, and the system that is represented by that model. It is also known as model (or simulator) bias, model error, or structural error. A "best input" approach typically defines discrepancy as the difference between the modelled system and the simulator when run at an input where output from the simulator conveys all it can about the system (see, e.g., Goldstein and Rougier, 2009). A practical definition from Williamson et al. (2014) is that "a climate model bias [simulator discrepancy] represents a structural error if that bias cannot be removed by changing the parameters without introducing more serious biases to the model". One of the main aims of the model development process is to efficiently identify important simulator discrepancies and correct them, or allow them to be taken into account in analyses – for example, during prediction using the simulator (e.g. Sexton et al., 2011).

Simulator discrepancy might be known ahead of time: perhaps a parameterisation of a process occurring at too high a resolution to simulate has a predictable effect on simulator behaviour. Alternatively, the discrepancy might be due to some missing and unknown process in the simulator, or to unknown parameterisation values. This might appear as a bias, only becoming apparent when output from the simulator is compared with observations of the real system. In both cases, the modeller must have a strategy for dealing with the discrepancy when using the simulator to make judgements about the system.

Simulator discrepancy is a major challenge during calibration. Kennedy and O'Hagan (2001) introduced a Bayesian framework for the task of the calibration of computationally expensive simulators. They urge the specification of a priori estimates of simulator discrepancy and offer methods to learn about that discrepancy by comparison of the simulator and observations. Failure to take simulator discrepancy into account in calibration can lead to overconfident and inaccurate estimates of the parameters and, consequently, the predictions of the simulator (e.g. Higdon et al., 2008; Brynjarsdóttir and O'Hagan, 2014). Often, there is an indeterminacy between parameter error and simulator discrepancy; that is, should we choose a different set of parameters as representing the "best" or should we add a simulator discrepancy term? Brynjarsdóttir and O'Hagan (2014) point out that strong prior information is required to distinguish between parameter uncertainty and discrepancy, and that this information is often lacking. Further, even inadequate (as opposed to outright wrong) specification of a simulator discrepancy can lead to overconfidence and bias in parameters and predictions.

1.2 Calibration of land surface components

Parametric uncertainty in the land surface and carbon cycle component of models is expected to represent a large fraction of current uncertainty in future climate projections (Booth et al., 2012, 2013; Huntingford et al., 2009). These components have been introduced into climate simulators more recently, and have not yet been subject to the depth of systematic evaluation as, for example, atmospheric components. There is much focus, therefore, on identifying parameter sets consistent with observed climate metrics and reducing future land carbon cycle uncertainty by identifying parts of simulator parameter space inconsistent with observed properties of the real climate system.

Using statistical and data assimilation approaches to constrain land surface simulator process parameters extends back at least to Sellers et al. (1996). Recent examples are community efforts to develop a systematic set of observations to benchmark land surface processes against metrics of real-world processes, for example the International Land Model Benchmarking Project (Luo et al., 2012) and PALS (Abramowitz, 2012). Such benchmarks use an extensive set of metrics, covering a broad cross section of simulator processes, enabling an assessment of overall skill and highlighting areas where simulators fall short. They provide a framework to assess improvements in skill arising from continual simulator development as well as prioritising resources towards processes that are less well simulated. Using many observed metrics for diverse processes also discourages overtuning to a particular process, to the detriment of wider simulator performance. One limitation of the benchmarking ap-

[1] Throughout the paper we often use *simulator* in place of "model", usually to distinguish an Earth system, climate, or other process model from a statistical model.

proach is that there is limited understanding of what information a given observed metric implies about the simulator formulation or parameters, or what this might imply about future projected changes.

1.3 Paper aims and outline

We aim to identify parameter sets of the land surface module of the climate simulator FAMOUS where simulator output and observations of forest fraction are consistent to an acceptable degree. An initial attempt using history matching suggests that FAMOUS is unable to simulate the Amazon forest and other forests simultaneously at any set of parameters within the experiment design. We argue that this is due to a fundamental simulator discrepancy, which has implications for constraining the input parameters of FAMOUS. We use a number of techniques to characterise and find the drivers of this structural error, before performing a second history match with an appropriate discrepancy function.

In Sect. 2 we describe the ensemble of a climate simulator, with the emulator and history-matching techniques used to explore simulator discrepancy described in Sects. 2.5 and 2.6 respectively. We perform an initial history-matching exercise in Sect. 3.1. We use the emulator to quantify relationships between the simulated forest fraction and a set of simulator input parameters in a sensitivity analysis in Sect. 3.2. Next, we measure the performance of the ensemble in simulating forest fraction in Sect. 3.3. We see how much input space would be ruled out as implausible in various scenarios of data combination and uncertainty budget in Sect. 3.4 and we learn what each individual observation tells us about input space in Sect. 3.5. In Sect. 3.6, we use the emulator and an implausibility measure to find the nominal “best” set of parameters for each forest and then project the consequences of using those parameters on the other forests. Finally, we perform a history-matching exercise with a credible discrepancy function to constrain input parameters in Sect. 3.7. In Sect. 4, we discuss the consequences of our findings for simulators of the Amazon rainforest before offering conclusions in Sect. 5.

2 Data and methods

2.1 The FAMOUS climate simulator

We use a pre-existing ensemble of the climate simulator FAMOUS throughout this study. The Fast Met Office UK Universities Simulator, FAMOUS (Jones et al., 2005; Smith et al., 2008), is a reduced-resolution climate simulator, based on, and tuned to replicate, the climate model HadCM3 (Gordon et al., 2000; Pope et al., 2000). Computational efficiency is gained primarily through reduced resolution. Atmospheric grid boxes are 4 times the size of HadCM3, and ocean grid boxes are also larger. There are fewer levels in the atmosphere (11 compared to 19), and the ocean time step is 12 h compared to 1 h for HadCM3. In the atmosphere, the time step is 1 h, doubled from HadCM3. The dynamic vegetation component is called TRIFFID and is described in detail in Cox (2001). FAMOUS runs approximately 10 times faster than HadCM3, making it ideal for running large ensembles, or long integrations, with modest supercomputing facilities.

Smith (2012) describes improvements to FAMOUS in sea ice, ozone, hydrological cycle conservation, and upper tropospheric dynamics. Williams et al. (2013) describe the inclusion of the carbon cycle in the simulator via perturbed physics ensembles of terrestrial and ocean parameters, of which the terrestrial ensemble is studied in this paper. Most recently, Williams et al. (2014) give details of inclusion of a scheme to simulate the cycling of oxygen in the ocean and its coupling with the carbon cycle.

The inclusion of vegetation in FAMOUS is documented in Williams et al. (2013), which introduces surface tiling in the newer MOSES2 scheme. Five different vegetation types are simulated: broadleaf and needleleaf trees, C_3 and C_4 grasses, and shrubs, each with a fractional coverage in a grid box. Several surface types represent the absence of vegetation: bare soil, land ice, urbanised land use, and inland water. Williams et al. (2013) describe the optimisation of carbon cycle parameters in the terrestrial and ocean domains, validated against observations and reanalysis products, and present climatologies using both fixed and dynamic vegetation.

2.2 Known biases in the climate of FAMOUS

FAMOUS shows a Northern Hemisphere-winter surface air temperature cold bias with respect to HadCM3 and also the overestimation of the fractions of needleleaf trees in North America and C_3 grassland in the northern part of Eurasia. The initial version of FAMOUS used the MOSES1 surface exchange scheme and did not explicitly describe the inclusion of any vegetation cover, instead using grid box averages of surface quantities such as root depth, surface albedo, and roughness length to describe momentum and water exchange between the surface and the atmosphere. Biases were already present in climate regimes (Gnanadesikan and Stouffer, 2006) relevant for the Amazon rainforest. Smith et al. (2008) noted that “the Amazon region is not wet enough for a fully humid region to exist”.

2.3 The ensemble

We use an ensemble of 100 simulations of FAMOUS detailed in Williams et al. (2013), and build upon the results of that study. The ensemble was run in order to test the utility of including the carbon cycle in enhancing the FAMOUS simulator. The ensemble design perturbs seven vegetation and land surface control parameters (see Table 1) in a Latin hypercube configuration (McKay et al., 1979). This kind of design efficiently spans parameter space, and is commonly used for constructing surface response type statistical models known as emulators (see, e.g., Urban and Fricker, 2010).

Table 1. Land surface input parameters for FAMOUS. PFT: plant functional type; LAI: leaf area index.

Parameter	Default	Units	Description
F0	0.875		Ratio of CO_2 concentrations inside and outside leaves at zero humidity deficit.
LAI_MIN	3		PFT must achieve this value of LAI before starting to contend with other PFTs for growing area.
NL0	0.03	kgN/kgC	Top leaf nitrogen concentration. The amount of nitrogen per amount of carbon.
R_GROW	0.250		Growth respiration fraction.
TUPP	36	°C	Control on variation of photosynthesis with temperature.
Q10	2		Control on soil respiration with temperature.
V_CRIT_ALPHA	0.5		Control of photosynthesis with soil moisture.

This design builds upon a previous ensemble run by Gregoire et al. (2010), and implicitly contains a further parameter, β, that indexes into that other ensemble. The β parameter indexes the top 10 performing simulations with regard to the atmospheric climate. The β parameter is uncorrelated with any land surface parameters and the simulator output, so we exclude it from the ensemble design, essentially treating it as a nuisance parameter.

Ranges for the land surface parameters follow those used in the study by Booth et al. (2012) and, as that paper makes clear, were chosen for a number of reasons, not necessarily to represent plausible ranges of their uncertainty. However, we are confident that the parameter ranges are wide enough to span the space which might a priori be considered reasonable.

The ensemble simulates the pre-industrial climate, with ensemble members spun up over a 200-year period to ensure that the vegetation is in equilibrium with the climate at 290 ppm of CO_2. The vegetation dynamics component of the simulator, TRIFFID, is run in "fast spin-up" mode, for the equivalent of 10 000 years for each decade of climate simulation, to allow for the long adjustment time of dynamic vegetation. The climatology is constructed using the final 30-year period of the ensemble.

2.4 Simulator outputs and observations

We compare simulated forest fraction against observations adapted from Loveland et al. (2000), consisting of regionally aggregated versions of the data used in the previous study by Williams et al. (2013). We use broadleaf only for the tropical forest, and a mixture of broadleaf and needleleaf for the North American forest. A spatial summary of the ensemble and observations can be found in Fig. 1. Figure 2 shows every input and summary output, plotted against each other. This shows the marginal relationships of the (1) inputs against the inputs (which, as expected, show no obvious relationship); (2) the strength of the marginal relationship between the inputs and outputs; and (3) the outputs against the outputs, which highlights where outputs vary together. Parameter ranges do not represent uncertainty, so the ensemble mean and standard deviation are not a meaningful representation of data uncertainty but provide a useful summary of the data. To summarise the forest fraction data, we find the mean forest fraction in each of the Amazon, Central African, South East Asian, North American, and global regions (see Fig. S1 in the Supplement for region details).

South East Asian and Central African forests vary together very strongly across the ensemble, whereas the Central African and North American forests show a weaker relationship. The latter might be expected, given the different structure of the North American forests, compared with the tropical. The scatter plot also identifies NL0 (leaf nitrogen) and V_CRIT_ALPHA (soil moisture control on photosynthesis) as being important controls on forest fraction, as the output seems to vary most with these parameters.

2.5 Training an emulator

FAMOUS is not fast enough to run at every point within input space required for our analyses. We therefore use a computationally cheap statistical proxy to the simulator, called an emulator. The emulator is a non-parametric regression model conditioned on the ensemble, providing a prediction of simulator output and corresponding uncertainty orders of magnitude faster than the original simulator. Once trained, any analysis that might have been done with the simulator can be done with the emulator, provided we include the extra uncertainty term to account for the fact that the emulator is not a perfect prediction of the simulator output. A useful introduction to emulators and their uses can be found in O'Hagan (2006), and recent developments in emulator use in climate studies can be found, for example, in Tran et al. (2016) and Bounceur et al. (2015).

We use a Gaussian process emulator that assumes zero uncertainty at points where the simulator has already been run, growing larger away from those points. We treat the output $g(x)$ of the simulator FAMOUS as a deterministic function of a vector of input parameters x. We train a number of emulators of the ensemble, the details for each depending on the application. Details of the emulator, training, and verification can be found in the Supplement.

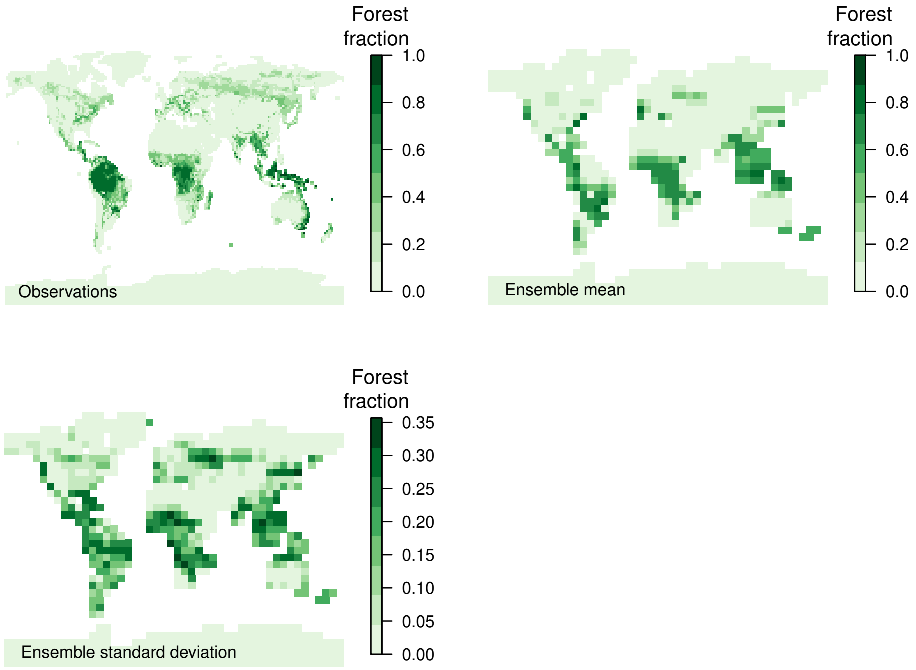

Figure 1. Observations of broadleaf forest fraction (top left panel). Mean (top right panel) and standard deviation (bottom left panel) of broadleaf forest fraction across the 100-member ensemble of FAMOUS.

2.6 History matching

After Williamson et al. (2014), we use history matching to find a region of parameter space consistent with observations to within the level of observational and acceptable simulator uncertainty. This requires finding a set of input parameters where the output of the simulator is tolerably close to the observations, given uncertainty in the observations and known deficiencies of the simulator. Constraining parameters in this way helps identify the range of projected futures of the forest consistent with the observations, rather than a single set of "best" parameters.

What distinguishes history matching from simulator calibration, where a probability distribution over the parameters is described, is that it rejects inputs inconsistent with observations, or otherwise classifies them as "not ruled out yet" (NROY). We regard NROY inputs as conditionally accepted, contingent on new observations or information. History matching was developed by Craig et al. (1997) and has been used extensively in hydrocarbon extraction sciences and astronomy (e.g. Vernon et al., 2010). Sometimes termed precalibration, it has been used to confront climate simulators with observations, for example by Lee et al. (2016), Williamson et al. (2013) and Holden et al. (2009). McNeall et al. (2013) investigated the potential of an observational dataset to constrain input space using history matching.

Observations of the system are denoted z, and we assume that they are made with uncorrelated and independent errors ϵ such that $z = y + \epsilon$, where y represents the true state of the climate being observed. Denoting the "best" set of input parameters x^*, and assuming the simulator contains a systematic structural error δ, the observations are related to input parameters

$$z = g\left(x^*\right) + \delta + \epsilon. \tag{1}$$

We could find the NROY region for x^* by running a large number of candidate points of the simulator in a Monte Carlo fashion. FAMOUS is not fast enough for this, and it is also our intention to develop methods that can be used on even more computationally expensive simulators. We therefore use the emulator as a proxy for the simulator output, replacing $g(x)$ with $\eta(x)$ in Eq. (1), and including a term for emulator uncertainty in the history-matching calculations.

Each candidate point is assigned an implausibility, I, according to the emulated forest fraction and uncertainty via Eq. (2). Inputs that produce forest fraction further from the observations are deemed more implausible. Those same inputs are less implausible if there is greater uncertainty about the observation, the simulator discrepancy, or the emulated output at that input:

$$I^2(x) = |z - E[\eta(x)]|^2 / [\mathrm{Var}(\eta(x)) + \mathrm{Var}(\delta) + \mathrm{Var}(\epsilon)]. \tag{2}$$

A threshold above which a candidate input can be safely ruled out as implausible is usually set to 3, roughly equivalent to a 95 % credible interval of a posterior distribution, if using a Bayesian analysis. This is due to Pukelsheim's three-sigma rule; that is, for any unimodal distribution, 95 % of the

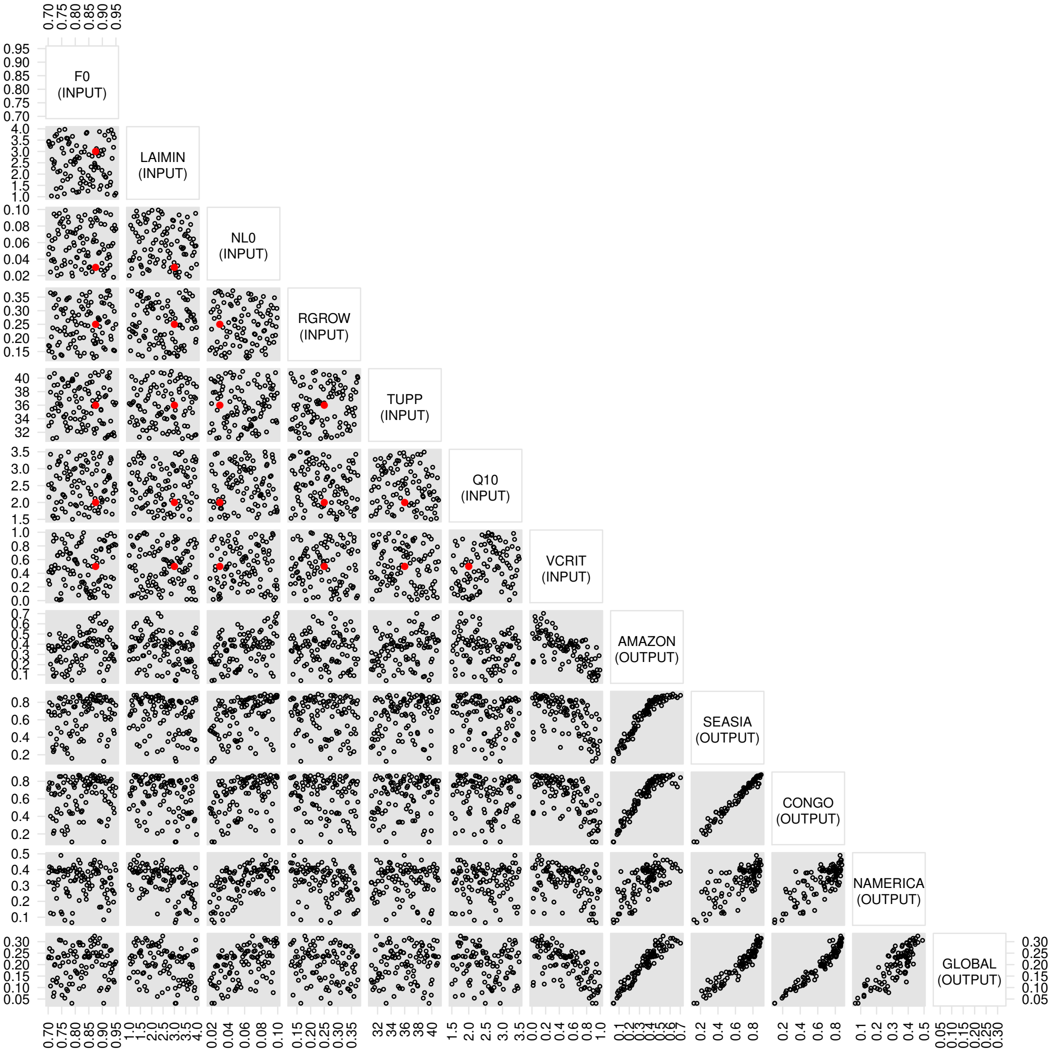

Figure 2. FAMOUS input parameters and forest fraction parameters, plotted against each other. Default inputs (not run) are marked in red.

probability mass will be contained within 3 standard deviations of the mean (Pukelsheim, 1994). Input parameter sets with an implausibility score below the threshold are designated NROY and retained for further analysis. This does not necessarily mean the input settings are *good*, merely that evidence from observations is not yet sufficient to rule them out as implausible. Inputs may be ruled out as more observations or simulator runs become available.

3 Analyses and results

3.1 An initial history match

In this section we find regions of land surface parameter space in FAMOUS that remain NROY given some defensible assumptions about observational uncertainty. Figure 3 shows how the regionally aggregated simulated forest fraction varies across the ensemble, compared with the corresponding observations. Although the simulator was not run with the "standard" parameter settings in the ensemble, we can use the emulator to estimate its output and uncertainty (±1 SD – standard deviation) at those settings; these are shown on the plot, in black.

The simulator run at the standard inputs significantly underestimates the forest fraction in the Amazon region, with a best estimate of > 0.3. The other tropical forests are slightly overestimated, North American forests are very slightly underestimated. Global forest fraction is simulated close to the observed fraction. Most ensemble members overestimate forest fraction in Central Africa, South East Asia, and North America. Some ensemble members simulate an Amazon forest fraction around, and above, the observed fraction. This gives us cause to hope that it is possible to find a set of parameters where the Amazon and other forests are simultaneously well simulated, without using a simulator discrepancy function.

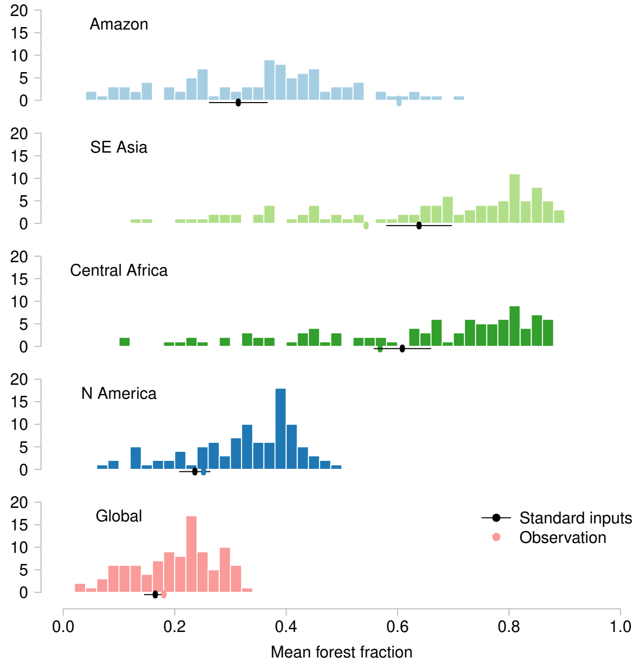

Figure 3. Histograms representing the number of ensemble members of a particular forest fraction in each region, as well as globally. Points plotted below the histograms represent the observed forest fraction (colours) and the forest fraction simulated at the "standard" parameters ±1 SD (black).

We aim to find regions of parameter space where simulator error is removed, or minimised to a level consistent with observational uncertainty. In practice, this requires finding a region where the large negative bias in Amazon forest fraction is minimised while keeping the other forests well represented.

On the advice of domain experts, we assume observational uncertainty of 0.05 (1 SD) in the Amazon, Central African, South East Asian, and North American forests as broadly representative, or at least usefully illustrative. This corresponds to an expectation that the true 95 % confidence interval is contained within the interval of ±0.15, following Pukelsheim's rule. This is nearly a third of the available range of zero to one, and it would be hard to argue that this represents an over-constraint.

We sample uniformly across input parameter space and run the emulator at these locations. We history-match the samples using all four individual forest observations and visualise the space where $\max[I] < 3$. Figure 4 shows a density pairs plot of the approximately 12 % of the 10 000 samples from the emulator that are not ruled out yet by the history match.

Does this region represent a viable set of inputs, perhaps to replace the default set of parameters, or should we include a non-zero discrepancy term (δ in Eq. 1)? Where it appears that we may have found regions where both Amazon and other forests are plausible, we are suspicious of this region, for three reasons. First, the default set of parameters is ruled out, in this case by comparison of the simulator with observations of the Amazon (Table 2). Second, it appears that in the active parameter space projections, these candidates are near the edges and corners of the input space considered plausible. The failure to rule out these points could be due to a relatively large emulator uncertainty, for example. Third, we plot the histograms of the "best estimate" emulator output at these NROY points (Fig. 5), and we see that they can be seen as *compromise candidates*. In general, if the simulator is run at points in this region, it will overestimate the Central African, South East Asian, and, most likely, North American forest fraction while underestimating the Amazon forest fraction. The candidate points are still included as NROY at these input values because of the combination of the emulator uncertainty and the assumed observational uncertainty.

Table 2. Implausibility I of forest observations at default input parameter setting of FAMOUS.

Observation	Implausibility I at default parameters
Amazon	3.99
Central Africa	0.56
South East Asia	1.24
North America	0.27

In the remainder of this section, we use a number of analysis techniques to investigate why a region on the edge of parameter space was initially considered plausible, and which does not contain the default parameter settings, is identified as NROY.

3.2 Finding the active parameters with sensitivity analysis

We perform a sensitivity analysis to identify the active subspace of simulator inputs and quantify relationships between inputs and outputs. In a descriptive sensitivity analysis, we show emulated mean regional and global forest fraction with inputs sampled from across input parameter space in a one-factor-at-a-time fashion, holding all but one parameter at their standard values while varying the remaining parameter (Fig. 6). The emulator is not a perfect representation of the simulator, and so we include the emulator uncertainty estimates at ±1 SD, shown as shaded regions in the plot.

V_CRIT_ALPHA, and NL0 are the most influential individual parameters and counter each other when both increased. The Q10 parameter has little or no influence on forest fraction. The TUPP parameter is important only to the Central African (termed "Congo" here, for brevity) and South East Asian forest fraction, much less important to the

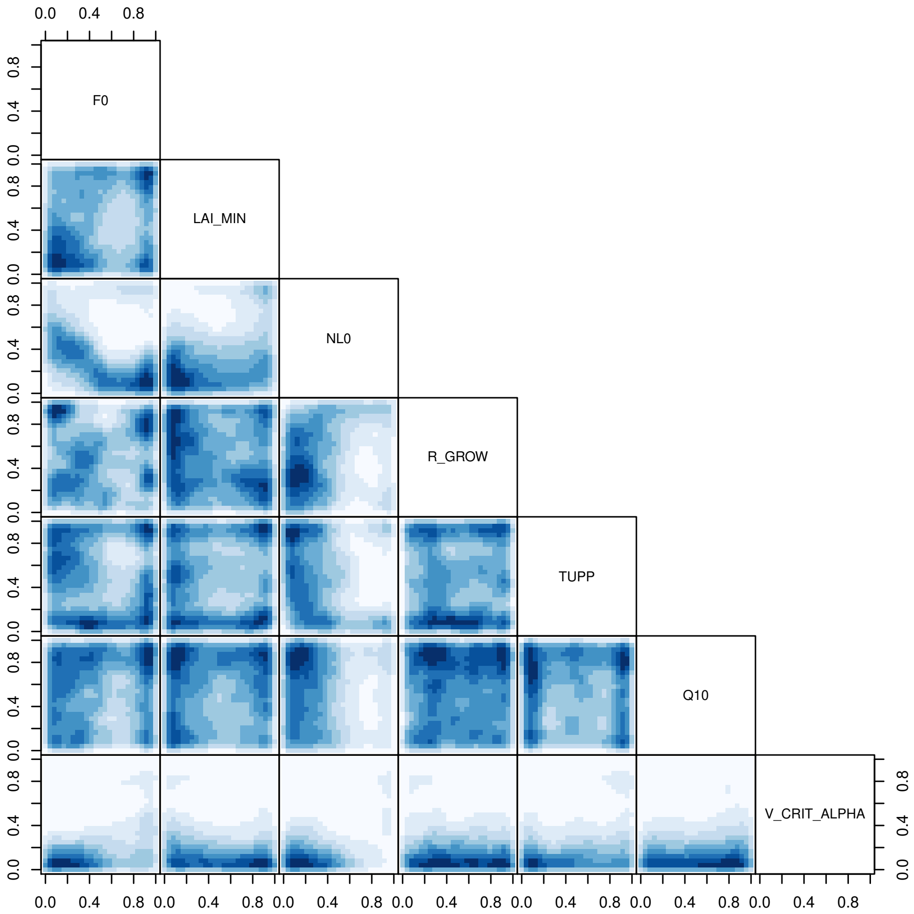

Figure 4. A density pairs plot of two-dimensional projections of parameter space. The blue areas represent the density of NROY points, using all of the data, with an assumed observational uncertainty of 0.05 (1 SD).

Amazon, and not important at all to the North American forests.

The relationships change across parameter space and are therefore dependent on the somewhat arbitrary range of the initial input parameters of the ensemble design. Sensitivity can change in importance as parts of input space are ruled out. For example, the forests are most sensitive to NL0 in the lower part of the ensemble range, and most sensitive to V_CRIT_ALPHA in the upper part of the ensemble range.

Following Carslaw et al. (2013), we quantify the sensitivity of the simulated forest fraction to the input parameters, using the FAST methodology (Saltelli et al., 1999), conveniently coded in the R package *sensitivity* (Pujol et al., 2015) and easily calculated using the emulator. We calculate the global sensitivity of the simulator output due to each input, as both a main effect and total effect, including interaction terms (Fig. 7). V_CRIT_ALPHA (soil moisture photosynthesis control parameter) is the most important parameter across the tropical forests and globally, with a total effect index of around 0.6. In tropical forests, NL0 (leaf nitrogen parameter) is next most important, with a total effect index between 0.2 and 0.3. In all cases, interaction terms are relatively unimportant, accounting for only a few percent of the variance. North American forests show slightly different results, with NL0 being the most important parameter with a sensitivity index near 0.4, followed by LAI_MIN (leaf area index parameter) at around 0.3 and V_CRIT_ALPHA at 0.25. This difference is unsurprising, as the North American forests are a mix of broadleaf and needleleaf trees, which will have different sensitivities from a broadleaf tropical forest.

Parameter Q10 has almost no influence on forest fraction, in line with the expectations of land surface modellers. This non-zero estimate of sensitivity is likely due to the fact that the emulator is not a perfect representation of the simulator, and a zero sensitivity is well within the uncertainty bounds of the sensitivity analysis. Parameters TUPP and R_GROW have very little impact on forest fraction. Parameter F0 has virtually no influence away from the tropics; conversely, LAI_MIN is only important in the North American forest.

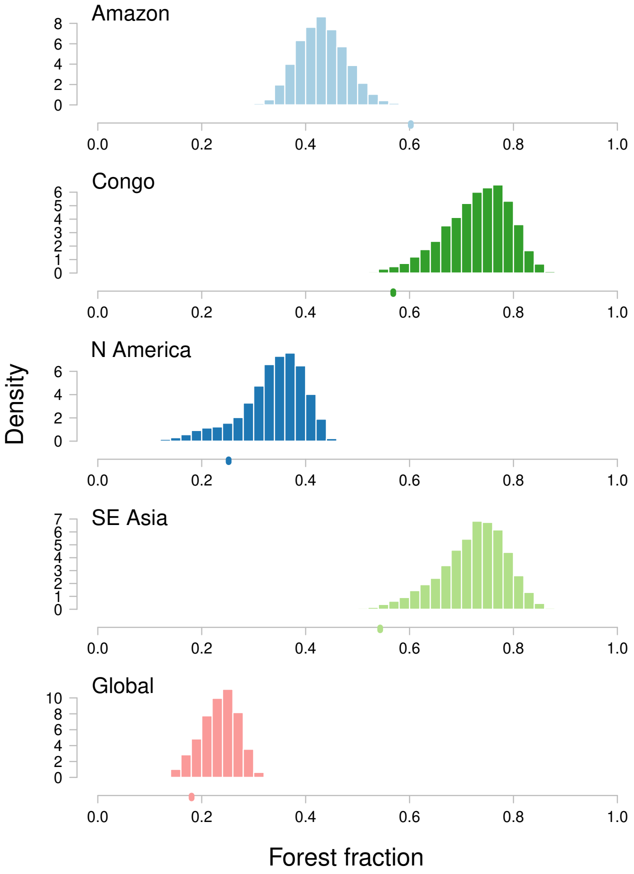

Figure 5. Best-estimate draws of forest fraction output from the emulator, at the set of points not ruled out yet when assuming a credible observational uncertainty. The value of the observed forest fractions is plotted as a single point on the corresponding x axes (a "rug plot").

3.3 Mapping simulator error in parameter space

In this section, we examine the ability of the simulator to reproduce the observed forest fraction, as well as how that ability varies across input parameter space, and assess the region of parameter space which is consistent with each of the forest fraction observations.

We show a map of simulator error in the two-dimensional space of the most important parameters identified in Sect. 3.2, in Fig. 8. We sample uniformly across all parameter space, and plot the mean emulated difference between simulator output and the observations for each point. The maps appear noisy because of the impact of randomly chosen values of the remaining dimensions, but the structure is clear. For the Central African, South East Asian, and North American forests there is a broad sweep of parameter space, running from low NL0, low V_CRIT_ALPHA to high NL0, high V_CRIT_ALPHA, where simulator error is close to zero. The Amazon input space does not have this region – only the high NL0, high V_CRIT_ALPHA corner has a simulator error close to zero, suggesting a bias not common to all forests. The lack of overlapping regions where simulator error is close to zero suggests we are unlikely to find a region where we do not need a simulator discrepancy term. It is possible to find a portion of parameter space where the error is similar for all simulator outputs in the low NL0, high V_CRIT_ALPHA corner. However, the error is rather large (at least −0.6) at this point.

3.4 How much input space is ruled out by combinations of observations?

We find the potential of the history-matching technique to rule out parameter space under a number of scenarios of tolerance to observational and simulator structural error. The denominator of Eq. (2) is the sum of the squared variances of the emulator, discrepancy, and observational uncertainty. Our emulator uncertainty is emergent, but we can experiment by assuming an overall uncertainty budget or by partitioning assumed uncertainty between observations and simulator discrepancy.

Different observations rule out different parts of parameter space, while combining observations can be a powerful method of ruling out large parts of parameter space. A number of approaches to combining data in history matching are discussed in Vernon et al. (2010) and Williamson et al. (2013). A simple strategy is to calculate max$[I]$ at a candidate input across all data independently and then reject those candidates with a value larger than 3 in any. A danger of this approach is that a single poorly specified emulator or simulator discrepancy term could lead to large swathes of parameter space being incorrectly ruled out. As the number of comparisons with data goes up, so does the probability of including a poorly specified simulator discrepancy. For example, comparing a simulator with a serious but undiagnosed bias could lead to all a priori plausible parameter space being ruled out as a poor match to the observations. It is important to first combine knowledge and judgement about the system being modelled, and the way that the parameters represent their real-world counterparts (or do not), before relying on observations to remove plausible parameter space.

A conservative approach is to reject a candidate point only if it is judged implausible using a number of measures. This will be more robust to a poorly specified simulator discrepancy term. Vernon et al. (2010) use the second and third highest implausibility score, where a simulator has implausibility scores for multiple outputs calculated. This is to guard against poor emulators, but in practice it works just as well for poorly specified simulator discrepancy. An alternative suggested by Vernon et al. (2010) is to use a multivariate measure of implausibility.

To understand the value of individual observations, we ask the following questions: what is our tolerance to error? And what level of uncertainty in observations or simulator discrepancy can we tolerate before our observations become in-

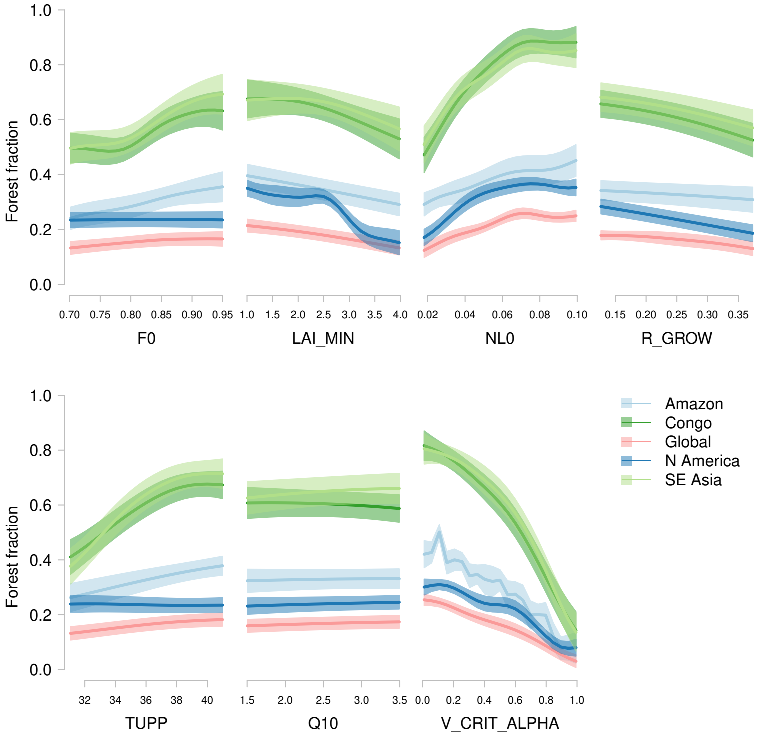

Figure 6. Marginal sensitivity of mean forest fraction to each input parameter in turn, with all other parameters held at standard values. Central lines represent the emulator mean, and shaded areas represent the estimate of emulator uncertainty, at the ±1 SD level.

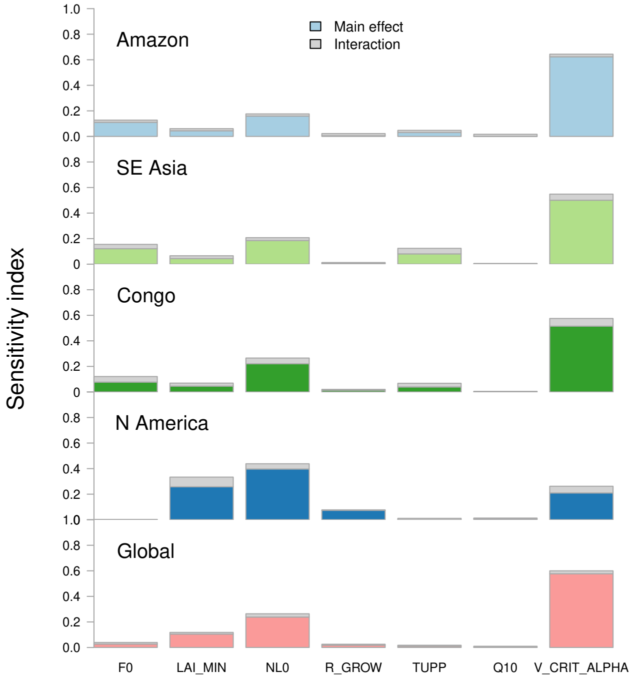

Figure 7. Sensitivity analysis of forest fraction via the FAST algorithm of Saltelli et al. (1999).

effective for history matching? Figure 9 shows the declining proportion of input parameter space ruled out as we increase tolerance to error in a number of scenarios. Tolerance to error is specified as a single standard deviation, so the full distribution of the uncertainty of the observation or discrepancy (e.g. the 95 % range) will be at least 3 times as large, using Pukelsheim's rule.

North American, South East Asian, and Central African forest observations constrain parameter space to between 40 and 50 % of parameter space, even when our tolerance to error is very low. The proportion of NROY space increases quickly, particularly using North American forest fraction, which becomes no constraint at all when our error tolerance is above 0.07 (1 SD). The other forests offer some constraint up to about 0.1 (1 SD), and the Amazon is more of a constraint, only losing power as a constraint when the standard deviation of our tolerance to error is above 0.15 (1 SD).

Combining data and using the maximum implausibility of any dataset improves the constraint, particularly when the tolerance to error is low. However, we urge caution. The fact that (a) the performance of the Amazon dataset appears different from the other observations and (b) that all parameter space is ruled out at lower values, even though there is emulator uncertainty, again raises concerns of a poorly specified Amazon simulator discrepancy.

A more robust calculation of tolerance to error can be found by excluding the Amazon observations and using the

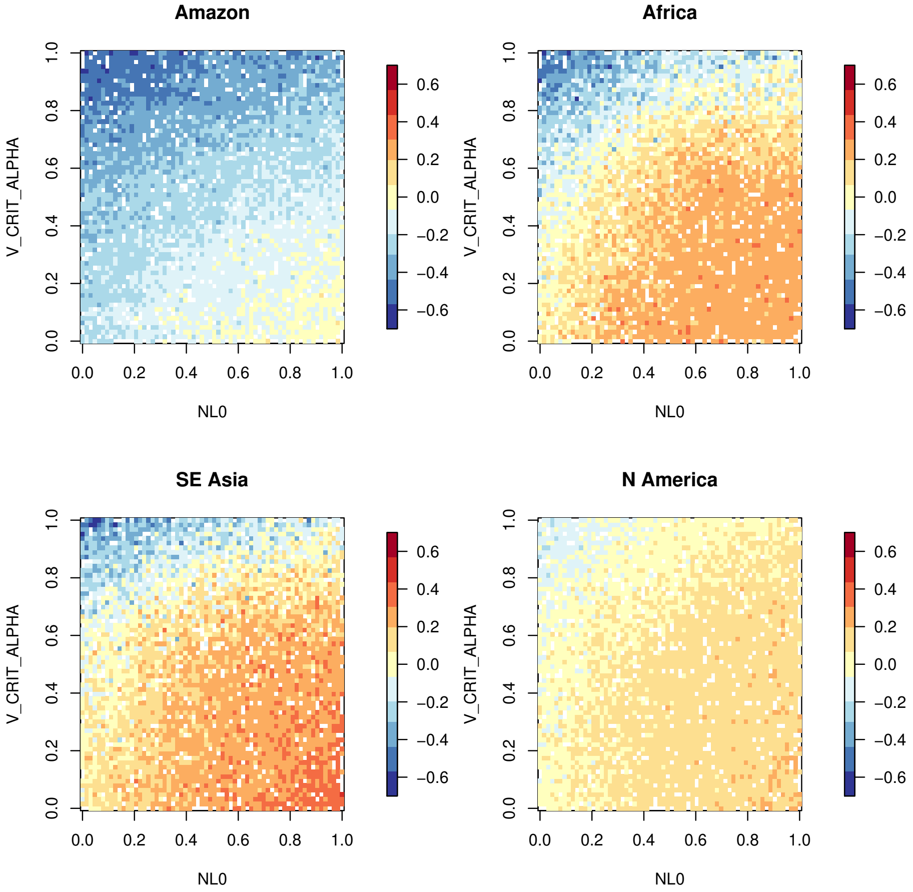

Figure 8. Maps of simulator error, in units of forest fraction, when projected into the two-dimensional space of the most active parameters, NL0 and V_CRIT_ALPHA.

maximum implausibility from the other observations. This excludes more input parameter space than any single observation on its own, up to a tolerance to error of around 0.85 (1 SD), where it performs in a similar manner to using South East Asian forest fraction.

To what extent do the input spaces that are NROY when history matching with two forests overlap? We suppose that data that suggest highly overlapping input spaces give us confidence that those input spaces are valid. Another perspective is that overlapping input spaces give us little extra information, and we should seek out those data that minimise overlap. We sample uniformly from the input space and test each point using a comparison with each forest observation to see if it is ruled out. If a point has the same status using both forests in the history match, we class that as an overlapping point. Table 3 gives the proportion of the samples which have the same status using each permutation of two forests for the history matching.

The most similar input space is found if we use the South East Asian and Central African rainforests. Comparing these forests with the North American forests gives a fairly high overlap – 61 and 66 % for South East Asia and Central Africa respectively. The Amazon has markedly lower overlap with the other forests: 40 % at the most with North America and only 26 % with South East Asia.

Table 3. Amount of overlap in NROY input space for forest combinations.

Forest A	Forest B	Input agreement (%)
Amazon	South East Asia	26
Amazon	Central Africa	33
Amazon	North America	40
South East Asia	Central Africa	84
South East Asia	North America	61
Central Africa	North America	66

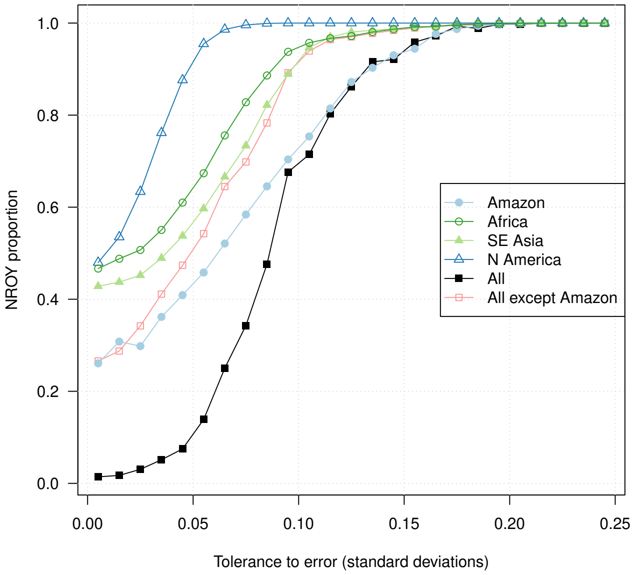

Figure 9. Proportion of NROY (not ruled out yet) input space plotted against "tolerance to error" – the total error budget including emulator, observational, and simulator discrepancy uncertainty.

3.5 What do the individual forests tell us about the best parameters?

To more fully explore the causes of simulator discrepancy and its consequences, we make the illustrative assumption that simulator discrepancy uncertainty is zero, and that observational uncertainty is very low. We sample a large number of points uniformly across input space and assume simulator discrepancy uncertainty of zero and an observational uncertainty of 0.01.

We classify as NROY only those emulated samples where the implausibility (or maximum implausibility in the case of combined data) is below 3. Setting such a demanding threshold allows us to find and describe the relatively small regions in input space where the simulator performs best, in two cases: first, using the South East Asian, Central Africa, and North American forest fraction in the history-matching exercise, and second using the Amazon forest fraction.

When plotted in two-dimensional projections (Fig. 10), the "best" set of parameters as defined by matching to the observed Amazon forest fraction, and to the other forests, form nearly non-overlapping sets in the most active subspace comprising V_CRIT_ALPHA and NL0. Again, we see a swathe of input parameter space, running from low V_CRIT_ALPHA, low NL0 through to high values of those parameters. This pattern is confirmed when using the individual datasets for history matching (not shown). The three non-Amazonian forests have a high degree of overlap of NROY space.

FAMOUS struggles to simulate both the Amazon and the other forests simultaneously, at any parameter combination when using a low threshold of implausibility. It is very difficult to reconcile the simulation of the Amazon simultaneously with the other forests if there is little uncertainty about the observations. A simulator discrepancy term and corresponding uncertainty is therefore necessary to attain an adequately performing simulator.

3.6 The forests at best parameters

To examine the implications of using each observation separately to tune the simulator, we use the emulator to project each forest at the set of "best" inputs: those where the simulator reproduces each forest with a very small tolerance of error. We then use the emulator to project the Amazon forest fraction using the "best" parameters for each forest, as well as the forest fraction for each of those forests using the "best" parameters for the Amazon in Fig. 11. As there is some uncertainty, due to emulator uncertainty and a small tolerance to error, these are plotted as histograms.

We find that the using the best set of parameters as defined for each non-Amazon forest would likely lead to an underestimate of the Amazon forest fraction by around 50 %, compared to the observed fraction (around 0.3, compared to an observation of around 0.6). Conversely, using the best parameters as defined for the Amazon leads to an overestimate of the other forests – around 0.3 for the tropical forests and 0.15 for the North American forest – even though the observed aggregate forest fraction is very similar for the tropical forests.

To further explore this difference, we project the "best" set of input parameters, found using the Amazon and African forest to match the simulator against, over a map of the entire FAMOUS land surface. In each case, an independent emulator is trained on the ensemble for each grid box. The maps of the mean forest fraction for each parameter set, and the difference between them, are shown in Fig. 12.

Even using the "best" Amazon parameters, the simulator underestimates the Amazon coverage in the north-east of South America. This makes it very difficult to simulate a sensible forest fraction, even when overestimating the forest fraction in places where the simulator does have forest cover.

3.7 History matching allowing for discrepancy in the Amazon

The previous sections show that the inputs where FAMOUS best simulates Central African, South East Asian, and North American forests cover a similar input space, whereas the best inputs for the Amazon are in a different region. A parsimonious approach would be to use a non-zero-mean discrepancy for the Amazon: allowing the Amazon to be less vigorous in our simulations, while maintaining that the simulator output should broadly match the other forests.

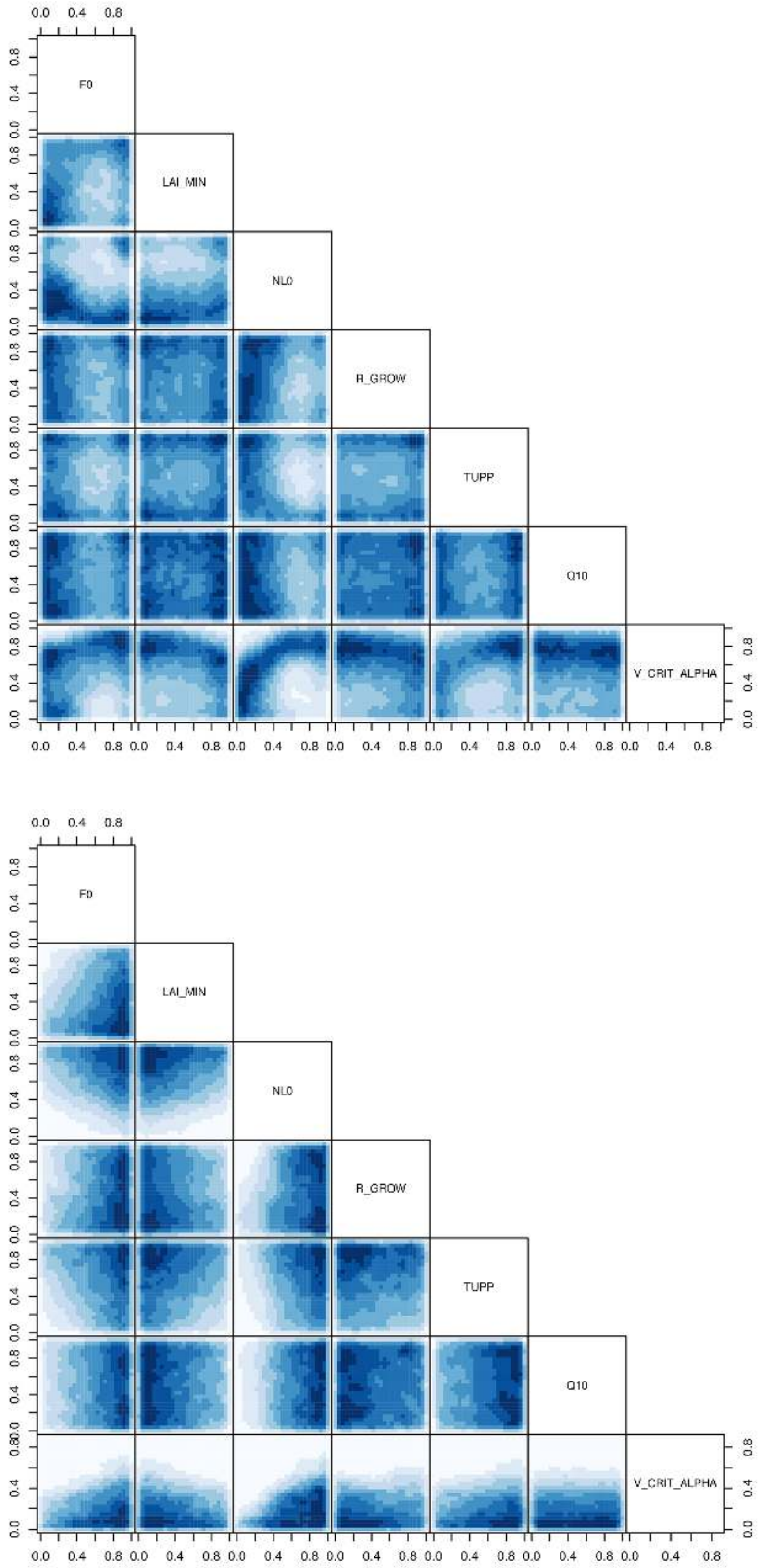

Figure 10. Marginal density of input parameter sets consistent with a very low "tolerance to error", as well as perfect observations, for the North American, South East Asian, and Central African forests combined (top panel) and the Amazon (bottom panel). Dark blue indicates those regions which have the highest concentration of NROY candidates and which are therefore most compatible with the observations.

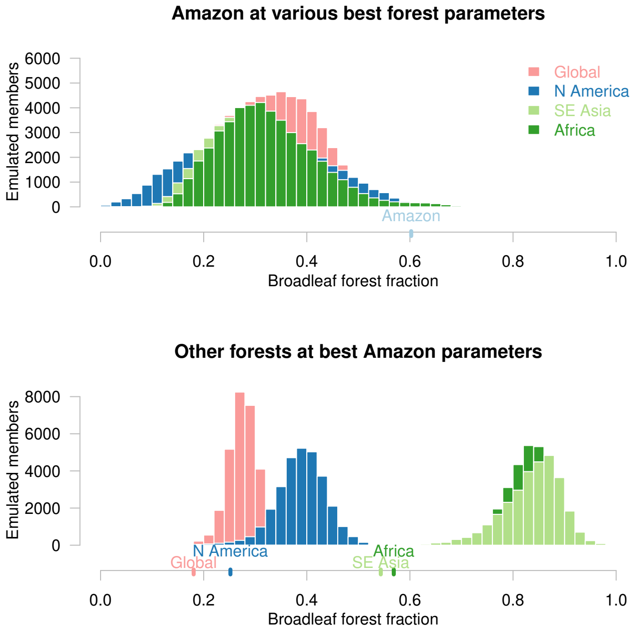

Figure 11. Top panel: forest fraction in the FAMOUS Amazon at the set of parameters where FAMOUS best matches each of the other forest observations. Bottom panel: other forests in FAMOUS at the set where the FAMOUS Amazon best matches observations. Observed forest fractions are shown as marks underneath the histograms.

We perform a history match using all of the forest observations, along with a simulator discrepancy term for the Amazon forest. We use the best estimate of the difference between Amazon observations and that simulated by FAMOUS at the default set of parameters as the best estimate of the discrepancy mean. The difference in forest fraction at the default parameters is approximately 0.3. Figure 13 shows the histograms of emulated simulator output using this discrepancy term, along with credible estimates for observational uncertainty (1 SD = 0.05) and tolerable discrepancy uncertainty (1 SD = 0.03). The corresponding two-dimensional density plots of NROY emulated input samples can be seen in Fig. 14. The remaining NROY input space represents around 57 % of the original input space defined by the input design, meaning that we have ruled out 43 % of the space. This contrasts with ruling out around 88 % of the space in the initial history match in Sect. 3.1. Marginal histograms of the relative density of NROY points for each individual input parameter (not shown) indicate that no part of the marginal input space is completely ruled out, and so we cannot "constrain" any of the parameters in an individual dimension.

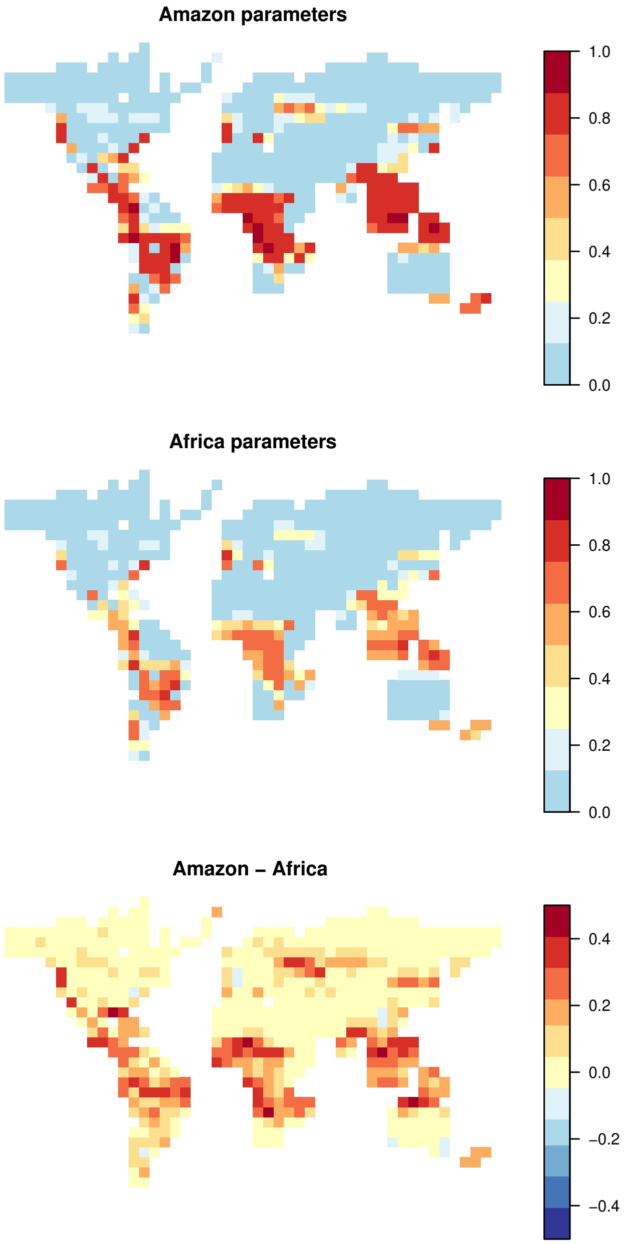

Figure 12. Maps of mean broadleaf forest fraction, over the "best" set of parameters found for the Amazon (top panel) and the Central African forest (centre panel). The difference between the two is mapped at the bottom panel.

4 Discussion

Our analysis illustrates the challenges in distinguishing between simulator discrepancy, parameter uncertainty, and observational uncertainty during simulator development. For example, forest fraction in the simulator can be tuned largely by using the two most active parameters: V_CRIT_ALPHA and NL0. As these parameters alter forest fraction in counteracting directions, a number of solutions can be found

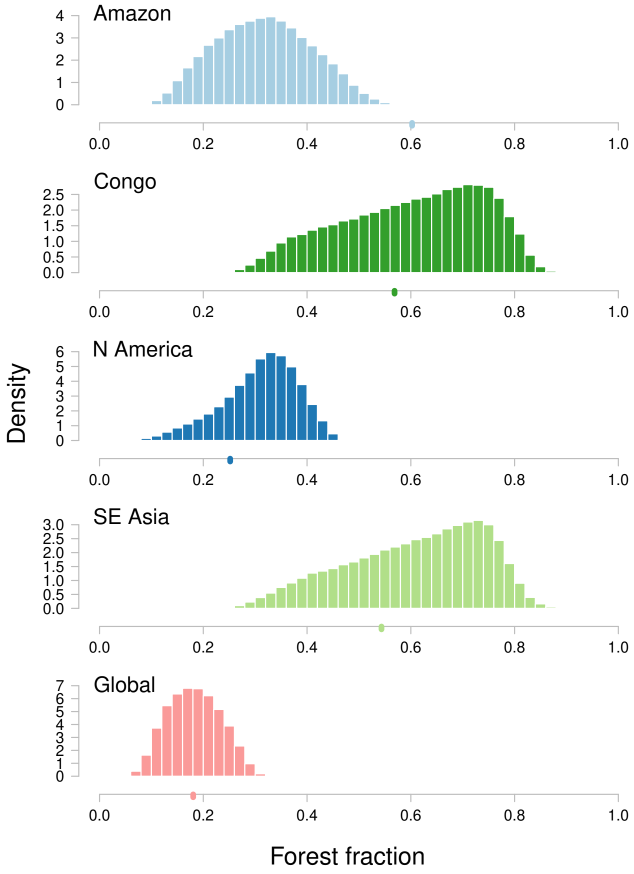

Figure 13. Histograms of emulated simulator output using credible estimates for observational uncertainty, a simulator discrepancy term for the Amazon, and credible discrepancy uncertainty.

that give plausible forest fractions. Information from outside sources about the "true" values of one these parameters might therefore offer a strong constraint on the value of the other. NL0 is the leaf nitrogen parameter – the ratio of nitrogen to carbon found in leaves. In theory, this is something that is well observed and recorded, but it is uncertain what value should be to reflect the observational range across the spatial scale of FAMOUS. Nitrogen content determines the maximum photosynthesis, and therefore how much CO_2 can be assimilated, or the productivity of a plant. Low (high) NL0 values correspond to low (high) nitrogen content and hence a low-productivity (high-productivity) plant. V_CRIT_ALPHA is the soil moisture threshold below which plants are water-limited, so if this parameter is high the plant is more often in a water-limited regime. If it is low, then a plant is not as often water-limited.

Using observations of the Amazon rainforest along with the other forests major forests in the history-matching exercise results in ruling out a large swathe of parameter space, including the default set of parameters, and leaving a corner of parameter space not ruled out yet. While it appears that here simulator output is tolerably close to the observations given a zero-mean discrepancy, there are good reasons to be suspicious of this region. For illustration, we imagine a situation where we are forced to choose between keeping the default parameters and including a simulator discrepancy function, or rejecting them and accepting a candidate from the new NROY region. Our choices will be dictated by the objective of our analysis: do we wish to provide only the best possible prediction, or do we wish to find parameter values which are, to some extent, "true"? For a simple prediction problem, we will be less concerned that the parameters more accurately reflect something we might measure in the real system, and might be less inclined to include a discrepancy term. However, sustainable development of the simulator requires that we get things right *for the right reason*. We argue that we should include a larger discrepancy function for the Amazon rather than ruling out the default parameters, for a number of reasons.

First, the NROY region excludes the default set of parameters, chosen as the result of multiple lines of evidence, scientific judgement, and experience using this and other simulators. Second, the NROY region is close to the edge of the ensemble in the active parameter subspace, so that emulator uncertainty, combined with the generous observational and discrepancy uncertainty, may dominate the implausibility calculation. Emulators tend to increase in uncertainty near the edge of an ensemble, as they are forced to extrapolate more than at the centre of the ensemble. Third, the information obtained from using each of the four forests shows that the Central African, South East Asian, and North American forests all indicate very similar, highly overlapping NROY regions. In contrast, the NROY region suggested by comparing FAMOUS to observations from the Amazon is very different. Finally, tuning to each of the "best" parameters for each of the forests suggests that the NROY region produces an inevitable compromise: the Amazon will be very likely be underestimated, and the other forests overestimated, if observational uncertainty is reduced. It is possible that there are correlated errors in the other forests, rather than in the Amazon. However, we argue that this is less likely, given that the other forests include tropical (like the Amazon) and the boreal forest of North America.

We therefore urge caution with a naive or automatic application of history-matching conclusions, particularly when using multiple observations for comparison with the simulator. Even in our relatively simple history-matching exercise, there is a clear need to include simulator discrepancy, to increase simulator discrepancy uncertainty, or to apply a conservative version of the measure of implausibility. One strategy, adopted, for example, by Vernon et al. (2014), is to reject parameter space that has a second- or third-highest implausibility metric larger than some threshold. This would be effective in the case of our comparison. Another strategy might be to reject only parameter space where the minimum implausibility is higher than some threshold. We believe that this would not rule out much input space in many circum-

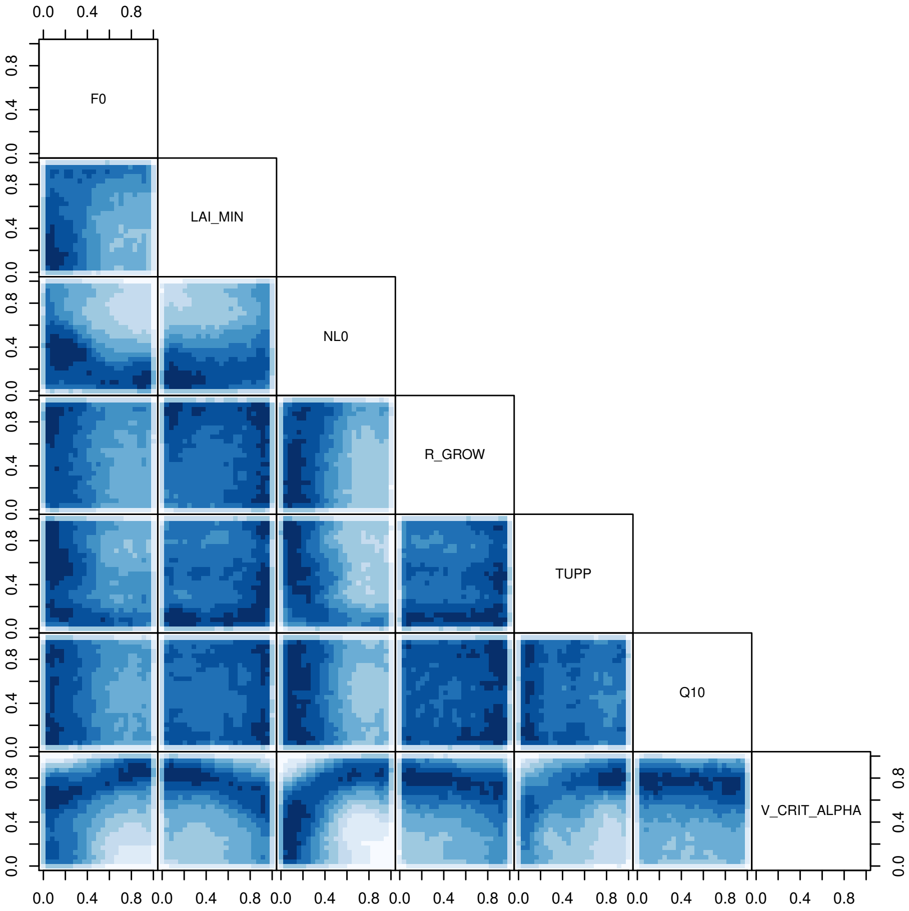

Figure 14. A density plot of the two-dimensional projections of NROY samples from the design input space, using all forest observations and a discrepancy function for the Amazon.

stances. We call for more research on the behaviour of measures of implausibility when the number of data comparisons is high and there is a chance that many of them may suffer from structural biases. Conducting a full probabilistic calibration as an alternative approach to our study might offer a powerful tool to overcome some of the difficulties we mention here. In particular, it would allow us to weight inputs as candidates for the "best", using the rules of probability, at the cost of expending effort in specifying prior distributions and likelihood functions.

We are able to offer a counter-example to the hypothesis of Williamson et al. (2014), who found regions of parameter space where what was thought to be a structural error in the simulator was significantly reduced. In this case, we believe it likely that better observations would simply confirm that the "best" regions of parameter space for the Amazon and other forests were non-overlapping. While individual forest fraction observations may have some uncertainty, we would expect the uncertainty on the differences between those observations to be smaller. A systematic bias in the way that the forests are measured would be common to all observations, for example, even though it would need to be taken into account in the uncertainty calculation for an individual observation.

We find that forest fraction does not offer a marginal constraint on the parameters: that is, there is little or no constraint on each parameter individually, but there is a significant constraint on the joint input space of the parameters. Approximately 43 % of a priori parameter space is ruled out, which is relatively little compared to other studies. This is explained by several factors: (1) the ensemble covers a relatively small input space, compared to other studies, due to the fact that the simulator is based on a well-studied climate model, HadCM3; (2) our observational uncertainty is assumed conservatively large; and (3) we have only a single wave of history matching. A further experiment could run the climate simulator within the NROY space in order to reduce emulator uncertainty and provide a basis to further rule out input space. The value of further waves of history matching might be diminished by the fact that the simulator likely has a large discrepancy in the Amazon, and the simulator discrepancy uncertainty is likely a large component of the overall uncertainty budget.

Causes of discrepancy

We suggest three possible causes of fundamental structural error which are *external* and *internal* to the vegetation model, although a combination of these causes is not ruled out. First, is there a problem with the emulator that would lead us to think that such a discrepancy exists? We believe that this is not the case, as the emulator performs sufficiently well across parameter space in cross-validation experiments (see Fig. S2).

Second, is there a missing processes in the vegetation model that impacts the Amazon or other forests in FAMOUS, or has the Amazon perhaps developed in other ways not seen in the other forests? For example, it is possible that the real Amazon can access water to a deeper level than other forests, through deep rooting. This would cause a *low Amazon* bias, seen in the simulator output. If the simulated Amazon cannot access water through deep enough roots, and simulator parameters were tuned to make Amazon as vigorous as in the real world, other forests would be more vigorous in the simulator than in observations. A bias that leads to a reduction in Amazon forest extent (such as that climatological or root depth) is likely to lead to further rainfall reductions, and its associated warming, as the region loses water cycling capability that the forest canopy provided. This is a feedback, and it can be expected to enhance any dry/warm bias that results from other factors and in turn enhance any forest loss. Such a simulator discrepancy could be countered by allowing different parameters in different regions, perhaps through ancillary parameter maps. Alternatively, the number of plant functional types allowed in the simulator could be increased – an approach adopted by many vegetation modelling efforts.

Finally, does the simulator simulate the climatic boundary conditions of the forest well enough? Malhi et al. (2009) and Staver et al. (2011) note the dramatic influence of climate on Amazon forest cover, albeit mediated by fire, a process not included in FAMOUS. Evidence from previous studies shows that HadCM3, which FAMOUS is designed to replicate, does have some climatic biases in the Amazon. Cox et al. (2004) find that rainfall in the Amazon is underestimated, particularly along the north-east coastline. Precipitation is underestimated by approximately 20 %. The dry season is too long (it starts a month early), and there is an underestimate of wet season rainfall. This precipitation anomaly persists in FAMOUS, although it is perhaps not as severe as in HadCM3 (Jones et al., 2005, Fig. 4). Good et al. (2008) note that simulated Amazon dry season precipitation is closely tied to meridional sea surface temperature gradients in the region. Joetzjer et al. (2013) and Yin et al. (2012) note similar climatic biases across the CMIP5 archive. We suggest that attributing the simulator discrepancy to these causes might be a fruitful direction for further study.

5 Conclusions

We analyse an ensemble of the fast climate simulator FAMOUS with the aim of constraining carbon cycle parameters through a comparison of simulator output with forest observations. We find that we are unable to constrain the parameters individually, but that areas of joint parameter space are effectively ruled out. With a defensible simulator discrepancy term for the Amazon and assumed observational uncertainty we are able to rule out 43 % of the input parameter space defined by the ensemble design.

We identify moisture control on photosynthesis (V_CRIT_ALPHA) as the most important parameter control on forest fraction, with the next most important parameter, leaf nitrogen (NL0), being approximately half as important, but still twice as important as any other parameter. These parameters have counteracting effects on the forest fraction, so we are unable to rule out a broad swathe of the joint space of these two parameters.

We suggest that we should exercise care if using observations of the Amazon rainforest to constrain the input parameters of FAMOUS, as an apparent structural bias in the climate simulator could lead to misleading results. Using the Amazon forest as an observational constraint suggests very different parts of input parameter space as *not implausible* compared to using other forests. Although we are able to find a region of parameter space that we are unable to rule out, given a defensible assumed observational uncertainty, we have reason to suspect that this region does not offer a credible alternative to default parameter settings. Further investigation reveals that choosing the region would systematically overestimate the forest fraction of the Central African, South East Asian, and North American forests, while simultaneously underestimating the Amazon. We fail to find a set of parameters that eliminates the discrepancy between the simulated fraction of the Amazon and other tropical and boreal forests. We suggest that we cannot find a set of vegetation model parameters that improve the Amazon without making the other forests worse. This satisfies the criterion of Williamson et al. (2014) to identify a simulator bias.

Using a history-matching technique, we investigate the limits of observational and simulator discrepancy uncertainty, beyond which observations no longer offer a constraint on input parameter space. We find that if this total error budget is larger than approximately 0.1 (1 SD of forest fraction), and excluding the Amazon rainforest as a comparison, the observations will not offer any form of constraint on the current ensemble, even in joint parameter space.

6 Data availability

Underlying data are available as an R data file: https://dx.doi.org/10.6084/m9.figshare.4244561.v1.

The Supplement related to this article is available online at doi:10.5194/esd-7-917-2016-supplement.

Author contributions. Doug McNeall and all authors designed the analysis. Doug McNeall conducted the analysis and wrote the paper. Jonny Williams provided the FAMOUS ensemble and Ben Booth provided the observed forest fraction data.

Competing interests. Richard Betts is a member of the editorial board of *Earth System Dynamics*.

Acknowledgements. This work was supported by the Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101). Doug McNeall was supported on secondment to Exeter University by the Met Office Academic Partnership (MOAP) for part of the work. Jonny Williams was supported by funding from Statoil ASA, Norway.

Edited by: J. Annan
Reviewed by: R. D. Wilkinson and one anonymous referee

References

Abramowitz, G.: Towards a public, standardized, diagnostic benchmarking system for land surface models, Geosci. Model Dev., 5, 819–827, doi:10.5194/gmd-5-819-2012, 2012.

Booth, B. B. B., Jones, C. D., Collins, M., Totterdell, I. J., Cox, P. M., Sitch, S., Huntingford, C., Betts, R. A., Harris, G. R., and Lloyd, J.: High sensitivity of future global warming to land carbon cycle processes, Environ. Res. Lett., 7, 024002, doi:10.1088/1748-9326/7/2/024002, 2012.

Booth, B. B. B., Bernie, D., McNeall, D., Hawkins, E., Caesar, J., Boulton, C., Friedlingstein, P., and Sexton, D. M. H.: Scenario and modelling uncertainty in global mean temperature change derived from emission-driven global climate models, Earth Syst. Dynam., 4, 95–108, doi:10.5194/esd-4-95-2013, 2013.

Bounceur, N., Crucifix, M., and Wilkinson, R. D.: Global sensitivity analysis of the climate–vegetation system to astronomical forcing: an emulator-based approach, Earth Syst. Dynam., 6, 205–224, doi:10.5194/esd-6-205-2015, 2015.

Brynjarsdóttir, J. and O'Hagan, A.: Learning about physical parameters: the importance of model discrepancy, Inverse Problems, 30, 114007, doi:10.1088/0266-5611/30/11/114007, 2014.

Carslaw, K., Lee, L., Reddington, C., Pringle, K., Rap, A., Forster, P., Mann, G., Spracklen, D., Woodhouse, M., Regayre, L., and Pierce, J. R.: Large contribution of natural aerosols to uncertainty in indirect forcing, Nature, 503, 67–71, 2013.

Cox, M. P., Betts, A. R., Collins, M., Harris, P. P., Huntingford, C., and Jones, D. C.: Amazonian forest dieback under climate-carbon cycle projections for the 21st century, Theor. Appl. Climatol., 78, 137–156, doi:10.1007/s00704-004-0049-4, 2004.

Cox, P. M.: Description of the TRIFFID dynamic global vegetation model, Tech. rep., Technical Note 24, Hadley Centre, United Kingdom Meteorological Office, Bracknell, UK, 2001.

Craig, P., Goldstein, M., Seheult, A., and Smith, J.: Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments, in: Case studies in Bayesian statistics, vol. 3, edited by: Gatsonis, C., Hodges, J., Kass, R., McCulloch, R., Rossi, P., and Singpurwalla, N., Springer-Verlag, New York, USA, 36–93, 1997.

Gnanadesikan, A. and Stouffer, R. J.: Diagnosing atmosphere-ocean general circulation model errors relevant to the terrestrial biosphere using the Koppen climate classification, Geophys. Res. Lett., 33, l22701, doi:10.1029/2006GL028098, 2006.

Goldstein, M. and Rougier, J.: Reified Bayesian modelling and inference for physical systems, J. Stat. Pl. Infer., 139, 1221–1239, 2009.

Good, P., Lowe, J. A., Collins, M., and Moufouma-Okia, W.: An objective tropical Atlantic sea surface temperature gradient index for studies of south Amazon dry-season climate variability and change, Philos. T. Roy. Soc. Lond. B, 363, 1761–1766, doi:10.1098/rstb.2007.0024, 2008.

Gordon, C., Cooper, C., Senior, A. C., Banks, H., Gregory, M. J., Johns, C. T., Mitchell, B. J. F., and Wood, A. R.: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments, Clim. Dynam., 16, 147–168, doi:10.1007/s003820050010, 2000.

Gregoire, L. J., Valdes, P. J., Payne, A. J., and Kahana, R.: Optimal tuning of a GCM using modern and glacial constraints, Clim. Dynam., 37, 705–719, doi:10.1007/s00382-010-0934-8, 2010.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M.: Computer Model Calibration Using High-Dimensional Output, J. Am. Stat. Assoc., 103, 570–583, doi:10.1198/016214507000000888, 2008.

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., and Wilkinson, R. D.: A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1, Clim. Dynam., 35, 785–806, doi:10.1007/s00382-009-0630-8, 2009.

Huntingford, C., Lowe, J. A., Booth, B. B. B., Jones, C. D., Harris, G. R., Gohar, L. K., and Meir, P.: Contributions of carbon cycle uncertainty to future climate projection spread, Tellus B, 61, 355–360, doi:10.1111/j.1600-0889.2009.00414.x, 2009.

Joetzjer, E., Douville, H., Delire, C., and Ciais, P.: Present-day and future Amazonian precipitation in global climate models: CMIP5 versus CMIP3, Clim. Dynam., 41, 2921–2936, doi:10.1007/s00382-012-1644-1, 2013.

Jones, C., Gregory, J., Thorpe, R., Cox, P., Murphy, J., Sexton, D., and Valdes, P.: Systematic optimisation and climate simulation of FAMOUS, a fast version of HadCM3, Clim. Dynam., 25, 189–204, doi:10.1007/s00382-005-0027-2, 2005.

Kennedy, M. and O'Hagan, A.: Bayesian calibration of computer models, J. Roy. Stat. Soc. B, 63, 425–464, 2001.

Lee, L. A., Reddington, C. L., and Carslaw, K. S.: On the relationship between aerosol model uncertainty and radiative forcing uncertainty, P. Natl. Acad. Sci. USA, 113, 5820–5827, doi:10.1073/pnas.1507050113, 2016.

Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L., and Merchant, J. W.: Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data, Int. J. Remote Sens., 21, 1303–1330, doi:10.1080/014311600210191, 2000.

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Mahecha, M., Niu, S. L., Norby, R., Piao, S. L., Qi, X., Peylin, P., Prentice, I. C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y. P., Xia, J. Y., Zaehle, S., and Zhou, X. H.: A framework for benchmarking land models, Biogeosciences, 9, 3857–3874, doi:10.5194/bg-9-3857-2012, 2012.

Malhi, Y., Aragão, L. E. O. C., Galbraith, D., Huntingford, C., Fisher, R., Zelazowski, P., Sitch, S., McSweeney, C., and Meir, P.: Exploring the likelihood and mechanism of a climate-change-induced dieback of the Amazon rainforest, P. Natl. Acad. Sci. USA, 106, 20610–20615, doi:10.1073/pnas.0804619106, 2009.

McKay, M., Beckman, R., and Conover, W.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 239–245, 1979.

McNeall, D. J., Challenor, P. G., Gattiker, J. R., and Stone, E. J.: The potential of an observational data set for calibration of a computationally expensive computer model, Geosci. Model Dev., 6, 1715–1728, doi:10.5194/gmd-6-1715-2013, 2013.

O'Hagan, A.: Bayesian analysis of computer code outputs: a tutorial, Reliab. Eng. Syst. Saf., 91, 1290–1300, 2006.

Pope, D. V., Gallani, L. M., Rowntree, R. P., and Stratton, A. R.: The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3, Clim. Dynam., 16, 123–146, doi:10.1007/s003820050009, 2000.

Pujol, G., Iooss, B., with contributions from Sebastien Da Veiga, A. J., Fruth, J., Gilquin, L., Guillaume, J., Gratiet, L. L., Lemaitre, P., Ramos, B., and Touati, T.: sensitivity: Sensitivity Analysis, r package version 1.11.1, https://CRAN.R-project.org/package=sensitivity, last access: 18 May 2015.

Pukelsheim, F.: The three sigma rule, Am. Stat., 48, 88–91, 1994.

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

Sellers, P., Randall, D., Collatz, G., Berry, J., Field, C., Dazlich, D., Zhang, C., Collelo, G., and Bounoua, L.: A Revised Land Surface Parameterization (SiB2) for Atmospheric GCMS. Part I: Model Formulation, J. Climate, 9, 676–705, doi:10.1175/1520-0442(1996)009<0676:ARLSPF>2.0.CO;2, 1996.

Sexton, D. M. H., Murphy, J. M., Collins, M., and Webb, M. J.: Multivariate probabilistic projections using imperfect climate models part I: outline of methodology, Clim. Dynam., 38, 2513–2542, doi:10.1007/s00382-011-1208-9, 2011.

Smith, R. S.: The FAMOUS climate model (versions XFXWB and XFHCC): description update to version XDBUA, Geosci. Model Dev., 5, 269–276, doi:10.5194/gmd-5-269-2012, 2012.

Smith, R. S., Gregory, J. M., and Osprey, A.: A description of the FAMOUS (version XDBUA) climate model and control run, Geosci. Model Dev., 1, 53–68, doi:10.5194/gmd-1-53-2008, 2008.

Staver, A. C., Archibald, S., and Levin, S. A.: The Global Extent and Determinants of Savanna and Forest as Alternative Biome States, Science, 334, 230–232, doi:10.1126/science.1210465, 2011.

Tran, G. T., Oliver, K. I., Toal, D. J., Holden, P. B., and Edwards, N. R.: Building a traceable climate model hierarchy with multi-level emulators, Adv. Stat. Clim. Meteorol. Oceanogr., 2, 17–37, doi:10.5194/ascmo-2-17-2016, 2016.

Urban, N. M. and Fricker, T. E.: A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth system model, Comput. Geosci., 36, 746–755, 2010.

Vernon, I., Goldstein, M., and Bower, R.: Galaxy formation: a Bayesian uncertainty analysis, Bayesian Anal., 5, 619–669, 2010.

Vernon, I., Goldstein, M., and Bower, R.: Galaxy Formation: Bayesian History Matching for the Observable Universe, Statist. Sci., 29, 81–90, doi:10.1214/12-STS412, 2014.

Williams, J. H. T., Smith, R. S., Valdes, P. J., Booth, B. B. B., and Osprey, A.: Optimising the FAMOUS climate model: inclusion of global carbon cycling, Geosci. Model Dev., 6, 141–160, doi:10.5194/gmd-6-141-2013, 2013.

Williams, J. H. T., Totterdell, I. J., Halloran, P. R., and Valdes, P. J.: Numerical simulations of oceanic oxygen cycling in the FAMOUS Earth-System model: FAMOUS-ES, version 1.0, Geosci. Model Dev., 7, 1419–1431, doi:10.5194/gmd-7-1419-2014, 2014.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Clim. Dynam., 41, 1703–1729, 2013.

Williamson, D., Blaker, A. T., Hampton, C., and Salter, J.: Identifying and removing structural biases in climate models with history matching, Clim. Dynam., 45, 1299–1324, doi:10.1007/s00382-014-2378-z, 2014.

Yin, L., Fu, R., Shevliakova, E., and Dickinson, R. E.: How well can CMIP5 simulate precipitation and its controlling processes over tropical South America?, Clim. Dynam., 41, 3127–3143, doi:10.1007/s00382-012-1582-y, 2012.