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)

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 ~~reduced resolution~~ global climate model ~~, FAMOUS. Using a history matching approach along with a~~ 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 ~~compare an ensemble of simulations of forest fraction with observations, and rule out~~ use history matching to rule out input parameter settings where the ~~forests are poorly simulated~~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. ~~Using observations of the Amazon forest to constrain input parameters leads to very different conclusions about the acceptable values of input parameters than using the other forests.~~

~~We characterise the structural model discrepancy, and explore the consequences of ignoring it in a history matching exercise. We use sensitivity analysis to find the parameters which have most impact on simulator error. We use the~~ We use the emulator to simulate the forest fraction at the best set of parameters implied by matching the model to the Amazon, ~~and to other major~~ 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 ~~calibrate the model~~ tune the simulator leads to a larger underestimate of the Amazon forest fraction.

~~Finally, we~~ We use sensitivity analysis to find the parameters which have 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 rule out just under half of joint parameter space ~~is ruled out~~ 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.

1 Introduction

~~A common practice in Earth system modelling is the parameterisation of processes which are too computationally expensive to represent explicitly . These parameterisations have associated numerical coefficients, quantitatively representing some~~ 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 ~~directly~~ represent a measurable physical quantity, or ~~they may~~ be a more abstract representation necessary due to the simplification of the modelled process. ~~There is often uncertainty about the~~ Uncertainty about the best value of the ~~parameter coefficients that should be used to best represent the system being simulated. It~~ coefficients means it may not be desirable ~~or practical~~ to choose a single value ~~of the coefficients~~ over all others~~, and uncertainty in the best choice of parameters~~ . 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 ~~appropriate values of these~~ parameterisation coefficients is a major research effort ~~that encompasses~~ encompassing domain specific, statistical and computational literature. ~~The coefficients~~ Coefficients are tuneable by ~~comparison of the behaviour of the~~ comparing the simulator with observations of the ~~real system, although there may also be direct measurements of the value of the coefficient or other~~ 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 ~~as an inverse problem in~~ in a formal probabilistic setting ~~, then it may also~~ it can provide probability distributions for the parameters of the ~~model~~simulator, and is known as *calibration*. ~~The process of choosing~~ Choosing a single best parameter set is ~~often called~~ *tuning*. *History matching* ~~provides a formal way of ruling~~ rules out parameter settings ~~that are inconsistent with observed data.~~

~~The motivation for calibration of a simulatoris twofold. First, a simulator which matches~~ 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 ~~predictions. Second, given an accurate simulator, a more tightly constrained parameter set should provide a narrower range of uncertainty in future~~ 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 ~~can also be~~ is also known as model (or simulator) bias, model error, or structural error. A ~~useful~~ '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. **?**). A practical definition from **?** 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

[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.

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. **?**).

Simulator discrepancy ~~is a major challenge during calibration. In many cases, 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? Sometimes, there is little or no information to distinguish between these two.~~

~~Simulator discrepancy~~ might be known ~~a priori - perhaps a computationally necessary simplification or parameterisation ,~~ ahead of time: perhaps a parameterisation of a process occurring at too high a resolution to simulate has a predictable effect on simulator ~~output~~behaviour. Alternatively, the discrepancy might be due to some missing and unknown process in the ~~model. This sort of discrepancy~~ simulator, or to unknown parameterisation values. This might appear as a bias, ~~and only become~~ only becoming apparent when output from the simulator is compared with observations of the ~~phenomena under study in 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. **?** introduced a Bayesian framework ~~to~~ 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 ~~model~~ simulator discrepancy into account in calibration can lead to overconfident and inaccurate estimates of the parameters, and consequently the predictions of the ~~model~~ simulator (e.g. ~~**?** , **?**)~~**??**). 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? **?** 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 (**?**, **?**, **?**). These components have been introduced into climate ~~models~~ 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, in identifying parameter sets ~~that are~~ consistent with observed climate metrics ~~, or at least~~ and reducing future land carbon cycle uncertainty by identifying ~~which parts of possible model parameter space are~~ parts of simulator parameter space inconsistent with observed properties of the real climate system.

~~There is also a long history of~~ Using statistical and data assimilation approaches ~~used to constrain process model parameters. In the land surface model context these extend back~~ to constrain land surface simulator process parameters extends back at least to **?**. 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 (**?**), and PALS (**?**). Such benchmarks ~~involve~~ use an extensive set of metrics, covering a broad cross-section of ~~model processes. These benchmarks enable~~ simulator processes, enabling an assessment of overall ~~model skill and highlight particular areas where the~~

model falls skill and highlighting areas where simulators fall short. They provide a useful framework to assess improvements in model skill that arise from continual model skill arising from continual simulator development as well as prioritising resources towards model processes that are less well simulated. Using a large number of many observed metrics for diverse aspects of the model processes also helps avoid model parameters being tuned to address processes also discourages overtuning to a particular process, to the detriment of wider model simulator performance. One of the limitations of the limitation of the benchmarking approach is that there is only limited current limited understanding of what information a given observed metric implies about the model simulator formulation or parameters, or what this might imply about future projected changes.

1.3 Paper aims and outline

Our aim is We aim to identify parameter sets for of the land surface module of the climate simulator FAMOUS where the simulator output and the 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. **??** we briefly describe the ensemble of a climate simulator, and describe the emulator and the history matching technique that we use history matching techniques used to explore simulator discrepancy in Sect. **??** and **??** respectively. We perform an initial history matching exercise in Sect. **??**. We use the emulator to quantify the relationships between the simulated forest fraction and a set of model simulator input parameters in a sensitivity analysis in Sect. **??**. Next, we measure the performance of the model ensemble in simulating forest fraction in Sect. **??**. We see how much input space would be ruled out as implausible in various scenarios of data combination and uncertainty budget in Sect. **??** and we learn what each individual observation tells us about input space in Sect. **??**. In Sect. **??**, we use the emulator and an implausibility measure to find the nominal "best" set of parameters for each forest, and 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. **??**. In Sect. **??**, we discuss the consequences of our findings for simulators of the Amazon rainforest . We offer before offering conclusions in Sect. **??**.

2 Data and Methods

2.1 The FAMOUS climate ~~model~~simulator

We use a pre-existing ensemble of the climate model simulator FAMOUS throughout this study. The Fast Met Office UK Universities Simulator FAMOUS (**??**) is a reduced resolution climate simulator, based on, and tuned to replicate, the climate model HadCM3 (**??**). Computational efficiency is gained primarily through reduced resolution. Atmospheric grid boxes are four times the size of HadCM3, and ocean gridboxes are also larger. There are fewer levels in the atmosphere (11 compared to

19), and the ocean timestep is 12 hours compared to 1 hour for HadCM3. In the atmosphere, the timestep is 1 hour, doubled from HadCM3. The dynamic vegetation component is called TRIFFID and is described in detail in **?**. FAMOUS runs approximately ten times faster than HadCM3, making it ideal for running large ensembles, or long integrations, with modest supercomputing facilities.

? describe improvements to FAMOUS in sea ice, ozone, hydrological cycle conservation and upper tropospheric dynamics. **?** describe the inclusion of the carbon cycle in the ~~model~~ simulator via perturbed physics ensembles of terrestrial and ocean parameters, of which the terrestrial ensemble is studied in this paper. Most recently, **?** give details of inclusion of a scheme to simulate the cycling of oxygen in the ocean and its coupling with the carbon cycle.

The ~~explicit~~ inclusion of vegetation in FAMOUS is documented in **?**, which introduces surface tiling in the newer MOSES2 scheme. Five different vegetation types are simulated: broadleaf and needleleaf trees, C3 and C4 grasses, and shrubs, each with a fractional coverage in a gridbox. Several surface types represent the absence of vegetation: bare soil, land ice, urbanised land use and inland water. **?** 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 C3 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 gridbox 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 (**?**) relevant for the Amazon rainforest. **?** noted: *"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 **?**, 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 ~~model~~simulator. The ensemble design perturbs 7 vegetation and land surface control parameters (see table **??**) in a latin hypercube configuration (**?**). This kind of design efficiently spans parameter space, and ~~has been shown to be better than others~~ is commonly used for constructing surface response type statistical models known as emulators ~~(?)~~ (see e.g. (**?**)).

This design builds upon a previous ensemble run by **?**, and implicitly contains a further parameter, β, that indexes into that other ensemble. The β parameter indexes the top 10 performing ~~models~~ simulations with regards to the atmospheric climate. The Beta parameter is uncorrelated with any land surface parameters and the ~~model~~ simulator output, so we exclude it from the ensemble design, essentially treating it as a nuisance parameter.

Table 1. Land surface input parameters for FAMOUS

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 ~~the leaf area index before it starts~~ LAI before starting to contend with othe
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.

Ranges for the land surface parameters follow those used in the study by **?**, 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 preindustrial 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 ~~use forest fraction as the primary simulator output for study. Observations of forest fraction were~~ compare simulated forest fraction against observations adapted from **?**, ~~and are~~ consisting of regionally aggregated versions of the data used in the previous study by **?**. 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 ~~forest fraction data in the ensemble~~ ensemble and observations can be found in ~~figures ??and ??~~ Fig. **??**. Figure **??** 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. ~~The former shows the spatial distribution of forest fraction in FAMOUS in maps of both the mean and standard deviation across the ensemble of 100 members.~~ Parameter ranges ~~are not explicitly chosen to~~ do not represent uncertainty, ~~and~~ 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 supplementary ~~matterial~~ material Fig. S1 for region details).

~~Figure ?? 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.~~

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~~, with more scatter. This~~ . 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

~~The simulator FAMOUS , although relatively computationally cheap,~~ FAMOUS is not fast enough to ~~evaluate at every viable candidate~~ run at every point within input space ~~, termed $\mathcal{X}$~~required for our analyses. We therefore use a computationally cheap statistical proxy to the simulator, called an emulator. The emulator ~~provides a~~ is a non-parameteric regression model conditioned on the ensemble, providing a prediction of simulator output ~~at any required untested input, many~~ 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, ~~with the proviso that we must include an~~ 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 **?** , and recent developments in emulator use in climate studies can be found, for example in **??** .

We use a ~~gaussian~~ Gaussian process emulator that assumes zero uncertainty at points where the ~~model~~ simulator has already been ~~evaluated~~run, growing larger away from those points~~, and dependent upon a set of hyperparameters that are trained at the same time as the emulator.~~

. We treat the output $g(x)$ of the simulator FAMOUS as a deterministic function of a vector of input parameters x. ~~The emulator is a nonlinear regression model conditioned on a sample, or ensemble, and provides a prediction of simulator output and corresponding uncertainty.~~

~~We build~~ We train a number of emulators of the ensemble, the details for each depending on the application. ~~All use the DiceKriging package (?) , in the statistical programming environment R (?) .~~ Details of the emulator, training and verification can be found in the supplementary material.

~~DiceKriging allows the user flexibility in specifying the emulator, and then estimates parameters of the statistical model using the training data. We verify the quality of the emulators, using a leave-one-out cross validation metric, ensuring that the accuracy and uncertainty estimates of the emulator are consistent across the ensemble (see supplementary materialFig. S2).~~

2.6 History matching

~~We aim to repeat the achievement of ? to~~ After **?** , we use history matching to find a region of parameter space ~~that is~~ consistent with observations to within the level of observational and acceptable simulator uncertainty. ~~In practice this means~~ This requires

finding a set of input parameters where the output of the ~~model is deemed~~ simulator is tolerably close to the observations, given uncertainty in the observations and known deficiencies of the ~~model~~simulator. Constraining parameters in this way ~~should help~~ helps identify the range of projected futures of the forest ~~that are~~ consistent with the observations, rather than a single set of "best" parameters.

A ~~key distinction from the practice of model calibration is that the set of statistically consistent inputs are not accepted, but instead are deemed~~ distinction from simulator calibration where a probability distribution over the parameters is described, history matching rejects inputs inconsistent with observations, or otherwise classifies them "Not Ruled Out Yet" (NROY). ~~As such, we regard them~~ We regard NROY inputs as conditionally accepted, contingent on new observations or information. History matching was developed by **?**, and has been used extensively in hydrocarbon extraction sciences, and astronomy (e.g. **?**). Sometimes termed precalibration, It has been used to confront climate simulators with observations, for example by **???**. **?** 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. ~~If we denote~~ Denoting the "best" ~~possible~~ set of input parameters x^*, and ~~assume that~~ assuming the simulator contains a systematic structural error δ, ~~then the observations can be related to~~ the observations are related to input parameters

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

~~If the simulator were fast enough to evaluate at~~ We could find the NROY region for x^* by running a large number of candidate points ~~for x^*, this region could be found by standard Monte Carlo or optimisation methods. Our simulator FAMOUS, although relatively computationally cheap,~~ of the simulator in a Monte Carlo fashion. FAMOUS is not fast enough for this~~.~~ ~~It~~, and it is also our intention to develop methods that can be used on even more computationally expensive simulators. We therefore ~~again~~ use the emulator as ~~an efficient~~ a proxy for the ~~model~~ simulator output, replacing $g(x)$ with $\eta(x)$ in Eq. (**??**), and including a term for emulator uncertainty in the history matching calculations.

Each ~~point in input space~~ candidate point is assigned an Implausibility I, according ~~to~~ according to the emulated forest fraction and uncertainty via Eq. (**??**). ~~The forest fraction at a sample of points in input space are calculated, along with uncertainties, using the emulator described above.~~ Inputs that produce forest fraction ~~that is~~ further from the observations are deemed more implausible. Those same inputs are less implausible if there is greater uncertainty about the observation, ~~about the model~~ 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 ~~of implausibility,~~ 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 for any unimodal distribution, 95% of the probability mass will be contained within 3 standard deviations of the mean (**?**).

~~Any input parameter set that has~~ Input parameter sets with an implausibility score below the threshold ~~is designated "Not Ruled Out Yet" (NROY), and is~~ are designated NROY and retained for further analysis. ~~It should be noted that this does not imply that the input setting is~~ This does not necessarily mean the input settings are *good* merely that ~~the~~ evidence from observations is not yet sufficient to rule ~~it~~ them out as implausible~~: this may change~~. Inputs may be ruled out as more observations ~~, or more~~ 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 **??** shows how the regionally aggregated simulated forest fraction varies across the ensemble~~. The figure shows histograms of the number of ensemble members with a particular forest fraction~~, compared with the corresponding observations. Although the simulator was not run at the "standard" ~~or "default"~~ parameter settings in the ensemble, we can use the emulator to estimate its output and uncertainty (± 1 standard deviation) at those settings, and show these on the plot, in black.

The ~~model~~ simulator run at the standard inputs significantly underestimates the forest fraction in the Amazon region~~by a considerable margin (a~~ , with a best estimate of ~~more than~~ $\geq$0.3~~)~~. The other tropical forests are slightly overestimated, ~~while the~~ North American forests are very slightly underestimated. Global forest fraction is simulated ~~very near~~ close to the observed fraction. Most ~~of the~~ ensemble members overestimate forest fraction in Central Africa, Southeast Asia, and North America. Some ensemble members simulate an Amazon forest fraction around, and ~~indeed~~ 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.

~~A target for a history matching exercise is to~~ We aim to 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.

~~We allow an~~ On the advice of domain experts, we assume observational uncertainty of 0.05 (one standard deviation) in ~~each of~~ 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 ~~range~~ is nearly a third of the available range of zero to one, and ~~we contend that~~ it would be hard to argue that this ~~is~~ represents an over-constraint.

We sample ~~from the emulator~~ uniformly across input parameter space ~~, history match~~ 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 **??** 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.

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
Southeast Asia	1.24
North America	0.27

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. **??**)? 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 **??**).

~~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 Southeast Asia 1.24 North America 0.27~~

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~~. When parameters near the edge of an experimental design are suggested as NROY by a simulator-data comparison, this can suggest an undiagnosed fundamental simulator discrepancy~~, for example. Third, we plot the histograms of the "best estimate" emulator output at these NROY points (Fig. **??**), 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. They are still included as NROY at these values because of the combination of the emulator uncertainty and the assumed observational uncertainty.

In the remainder of this section, we use a number of analysis techniques to investigate why a region on the edge of parameter space initially considered plausible, that 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 ~~model~~ simulator inputs and quantify relationships between ~~the model~~ 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. **??**). The emulator is not a perfect representation of the simulator, and so we include the emulator uncertainty estimates at $\pm$ one standard deviation, shown as shaded regions in the plot.

V_CRIT_ALPHA, and NL0 are the most influential individual parameters ~~when considered across the entire ensemble,~~ and counter each other when both ~~raised~~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 Southeast Asian forest fraction, much less important to the 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 (**?**), we quantify the sensitivity of the simulated forest fraction to the input parameters, using the FAST methodology (**?**), ~~as~~ conveniently coded in the R package *Sensitivity* (**?**), and easily calculated using the emulator. We calculate the global sensitivity of the ~~model~~ simulator output due to each input, as both a main effect and total effect, including interaction terms (Fig. **??**). 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 ~~an~~ 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 ~~expectation from~~ the expectations of land surface modellers. ~~The~~ This non-zero estimate of sensitivity ~~here is very~~ 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. ~~If we sum the total effects (first order plus interactions) for all of the forests types excluding global (as it is largely made up of our forests), we obtain a rank for each parameter (table **??**).~~

~~Total effect sum in sensitivity analysis.~~

~~Parameter Total effect sum VSUBSCRIPTNBCRITSUBSCRIPTNBALPHA 2.03 NL0 1.09 LAISUBSCRIPTNBMIN 0.53 F0 0.41 TUPP 0.25 RSUBSCRIPTNBGROW 0.13 Q10 0.04~~

3.3 Mapping simulator error in parameter space

In this section, we examine the ability of the simulator to reproduce the observed forest fraction, 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 the two dimensional space of the most important parameters identified in Sect. **??**~~parameter space~~, in Fig. **??**. We sample ~~the emulator~~ uniformly across all parameter space, and plot the mean ~~predicted difference between model~~ 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, Southeast 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 ~~bias in the model that is~~ a bias not common to all ~~of the forests~~forests. The fact that the regions of this reduced input space where the simulator error is close to zero do not overlap, means we are more unlikely to find parameter sets where a simulator discrepancy term is not needed. It is possible to find a portion of parameter space where the ~~bias~~ error is similar for all simulator outputs in the low NL0, high V_CRIT_ALPHA corner. However, the ~~bias~~ error is rather large (at least -0.6) at this point.

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

~~? discuss treating the model discrepancy as a "tolerance to error". We take this approach, and~~ We find the potential of the history matching technique to rule out parameter space under a number of scenarios of ~~observational and tolerance to model~~ tolerance to observational and simulator structural error. The denominator of Eq. (**??**) is the sum of the squared variances of the emulator, discrepancy, and observational uncertainty. Our emulator uncertainty is ~~set~~emergent, but we can experiment ~~with the~~ by assuming an overall uncertainty budget~~by partitioning~~ , or by partitioning assumed uncertainty between observations and ~~model discrepancyat will~~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 **?** and **?**. A simple strategy is to calculate $max[I]$ at a candidate input across all data independently, and reject those candidates with a value larger than 3 in any. A danger of ~~history matching using $max[I]$~~ this approach is that a single poorly specified emulator or ~~model~~ 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 ~~model~~ simulator discrepancy. For example, comparing a ~~model~~ 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. ~~For that reason, it~~ 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 don't), before ~~simply~~ relying on observations to remove plausible parameter space.

A conservative ~~measure~~ approach is to reject a candidate point only if it is judged implausible using a number of measures. This will ~~tend to~~ be more robust to a poorly specified ~~model~~ simulator discrepancy term. **?** use the 2nd and 3rd highest implausibility score, where a simulator has implausibility scores for multiple outputs calculated. This is to guard against poor emulators, but in practice works just as well for poorly specified ~~model~~ simulator discrepancy. An alternative suggested by **?** is to use a multivariate measure of implausibility.

To understand the value of individual observations, we ask *what is our tolerance to error?* What level of uncertainty in observations or ~~model discrepancy (or both)~~ simulator discrepancy can we tolerate before our observations become ineffective for history matching? Figure **??** shows the declining proportion of input parameter space ruled out as we increase ~~our~~ tolerance to error ~~,~~ in a number of scenarios. ~~Coloured lines indicate use of the individual, and combinations of, the forest fraction observations.~~ Tolerance to error is specified as a single standard deviation ~~, so in practice,~~ so the full distribution of the uncertainty of the observation or discrepancy (e.g. the 95% range) will be at least three times as large, using Pukelsheim's rule.

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

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

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 standard deviation). The other forests offer some constraint up to about 0.1 (1 standard deviation), and the Amazon is more of a constraint, only ~~completely~~ losing power as a constraint when the standard deviation of our tolerance to error is above 0.15 (1 standard deviation).

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

~~An alternative and perhaps~~ A more robust calculation of tolerance to error can ~~therefore~~ be found by excluding the Amazon observations and using the 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 standard deviation), where it performs in a similar manner to using Southeast 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 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~~or not~~. If a point has the same status using both forests in the history match, we class that as an overlapping point. Table **??** 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 Southeast Asian and Central African rainforests. Comparing these forests with the North American forests gives a fairly high overlap - 61% and 66% for Southeast 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.

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

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

We ~~keep~~ 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 ~~model~~ simulator performs best, in two cases. First, using the South East Asian, Central Africa and North American forest fraction in the history matching exercise, second using the Amazon forest fraction.

Plotted in two-dimensional projections in Fig. **??**, we see that the "best" set of parameters as defined by matching to the observed Amazon forest fraction, and to the other forests, form almost 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 high values of those parameters. This pattern is confirmed when using the individual data sets for history matching (not shown). The three non-Amazonian forests have a high degree of overlap of NROY space.

~~When run at a single parameter set,~~ FAMOUS struggles to simulate both the Amazon and the other forests simultaneously, at any parameter combination when using a low threshold of implausibility. ~~The implication of this is that it is~~ It is very difficult to reconcile the ~~model~~ simulation of the Amazon simultaneously with the other forests if there is little uncertainty about the observations. A ~~model~~ simulator discrepancy term and corresponding uncertainty ~~of some form is~~ is therefore necessary to attain an adequately performing simulator.

~~The emulator offers the advantage of flexibility, and we can predict the implausibility at any point in parameter space, identifying regions of input space where the model output is inconsistent with the observations. For example, in Fig. **??**, two parameters are varied across the full ensemble range, while all other parameters are held at their default value. The green point marks the default input value, projected into the two-dimensional space. For this illustrative example, we use a "tolerance to error" of 0.1 (1 standard deviation), which is the assumed sum of observation and discrepancy uncertainty.~~

~~Using the Central African (CONGO for brevity) rainforest to estimate implausibility of each point in parameter space, we see that the standard inputs are located in a deep "valley" of low implausibility. Generally, the implausibility is very low at the standard settings. There are regions where implausibility may be equally low or lower, existing as planes within the multidimensional space. However, there appears to be no evidence that the standard set is implausible, given this data.~~

~~In contrast, using the Amazon as an observation, the shape of the plausible regions seems very different when projected into this two dimensional space. There are no longer valleys of NROY space, but a larger region that appears off to one side of the design input space. In addition, the standard values are often close to or at the boundaries of implausible space.~~

3.6 The forests at best parameters

To examine the implications of using each observation separately to tune the ~~model~~simulator, we use the emulator to project the each forest at the set of "best" inputs~~for the alternative forests. We find input parameters where the model~~ : 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, and the forest fraction for each of those forests using the "best" parameters for the Amazon in Fig. **??**. 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 ~~most~~ 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 ~~. This occurs even even~~ - 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, is shown in Fig. **??**.

~~We see that even~~ 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 ~~approach~~ simulate a sensible forest fraction, even when ~~boosting~~ overestimating the forest fraction in places where the ~~model~~ simulator does have forest cover.

3.7 History matching allowing for discrepancy in the Amazon

~~Taken together, the analysis in the~~ 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. ~~This suggests that we should~~ 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. ~~We do not have enough information to create a more detailed discrepancy function: for example, one that varies across parameter space.~~

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 **??** shows the histograms of ~~NROY input space~~ emulated simulator output using this discrepancy term, along with credible estimates for observational uncertainty (1 standard deviation = 0.05) and tolerable discrepancy uncertainty (1 standard deviation = 0.03). The corresponding two-dimensional density plots of NROY emulated input samples can be seen in Fig. **??**. 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 inital history match in Sect. **??**.

~~Finally, marginal~~ Marginal histograms of the relative density of NROY points for each individual input parameter ~~are shownin Fig. ??. These~~ (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. ~~However, the relative frequency of NROY points is higher in some locations than others - low in NL0 and high in VSUBSCRIPTNBCRITSUBSCRIPTNBALPHA for example, suggesting a higher probability that the best estimates of the parameters is in these regions.~~

4 Discussion

~~Uncertainty in carbon cycle and land surface process contributes significantly to uncertainty in future climate change. There are a large number of uncertain input parameters to carbon cycle and land surface components of climate models, and our study attempts to use comparisons of the model with observed data to constrain some of the key parameters. 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 43of a priori parameter space is ruled out, which is relatively little compared to other studies. This is explained by two factors: 1) our observational uncertainty is assumed conservatively large, and 2) we have only a single wave of history matching. A further experiment could run the climate model 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 model discrepancy uncertainty is likely a large component of the overall uncertainty budget.~~

Our analysis illustrates the challenges in distinguishing between ~~model~~ simulator discrepancy, parameter uncertainty and observational uncertainty during ~~model~~ simulator development. For example, forest fraction in the ~~model~~ 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 that give plausible forest fractions. Information from outside sources about the "true" ~~(or appropriate)~~ 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 (high) productivity plant. V_CRIT_ALPHA is the soil moisture threshold below which plants are water limited~~. So~~ , so if this parameter is high ~~,~~ ~~then~~ the plant is more often in a water limited regime. ~~Where as if~~ If it is low, then a plant is not as often water limited.

~~If we use~~ Using observations of the Amazon rainforest ~~,~~ along with the other forests major forests in the history matching exercise ~~, we find that we can rule~~ results in ruling out a large swathe of parameter space, including the ~~standard~~ default set of parameters~~. There appears to be~~ , and leaving a corner of parameter space ~~that is NROY under these conditions~~Not Ruled Out

Yet. While it ~~first appears that we have found a region of parameter space where the model~~ 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. ~~The 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 ~~uncertainty~~ emulator uncertainty, combined with the generous observational and discrepancy uncertainty, may dominate the implausibility calculation. ~~The region excludes the default set of parameters, which are chosen the result of multiple lines of evidence and scientific judgement. Further~~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, Southeast 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. ~~Should we trust this region as NROY?~~

~~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 or candidates from the new NROY region . We argue that~~ 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 ~~fact that three of four data sets - and in different regions and types of forest - give us similar information about the parameters, and that they all include the default parameter settings as NROY, suggests that we should include a simulator discrepancy function~~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 ~~model~~ simulator discrepancy, or increase ~~model~~ simulator discrepancy uncertainty, or to apply a conservative version of the measure of implausibility. One strategy, adopted for example by **?** 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 circumstances. 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 **?**, who found regions of parameter space where what was thought a structural error in the ~~model~~ 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.

4.1 Causes of discrepancy

~~What could cause this fundamental structural error in the Amazon? There are~~ We suggest three possible causes of fundamental structural error - *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 ~~could cause such a bias~~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 supplementary material Fig. S2).

Second, is there a missing processes in the vegetation model, that impacts the Amazon ~~in FAMOUS? It~~ or other forests in FAMOUS, or perhaps has the Amazon has 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 ~~model~~ simulator output. If the simulated Amazon can't access water through deep enough roots, and ~~model~~ simulator parameters were tuned to make Amazon as vigorous as real world, other forests would be more vigorous in the ~~model~~ 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 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 ~~model~~ simulator simulate the climatic boundary conditions of the forest well enough? **?** and **?** note the dramatic influence of climate on Amazon forest cover, albeit mediated by fire~~and~~, 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. **?** 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 is perhaps not as severe as in HadCM3 (Jones et al. 2005, Fig. 4). **?** note that simulated Amazon dry season precipitation is closely tied to meridional sea surface temperature gradients in the region. **?** and **?** 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 ~~model~~ 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 ~~model~~ 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 leaf nitrogen (NL0), parameter being approximately half as important, and that 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 ~~model~~ 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* than 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 **?** to identify a simulator bias.

Using a history matching technique, we investigate the limits of observational and ~~model~~ 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 standard deviation 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.

Review comments and responses

Reviewer 1 (Anonymous, denoted R1:)

- R1: Review of paper: "The impact of structural error on parameter constraint in a climate model" by Doug McNeall et al. Thank you for inviting me to review this paper. The paper is interesting and important as it addresses whether a component of a GCM can be calibrated for one part of the globe, but applied elsewhere. Climate models are heavily dependent on transferability of parameterisation of sub-model structure, and a knowledge of when this fails is important. I can see the aim of the paper, and it will be useful to have in the literature. However there did seem to be a slightly excessive use of statistical terminology. That's fine if the statistics is of standard form, but that's not the case here as the methods utilised are more novel. Please ensure that the literature is cited sufficiently well that any part of this paper can be understood by calling upon the appropriate referenced papers.

- Response: With a paper at the interface of climate modelling and statistics, finding the correct balance of technical versus general description will always be difficult. Our strategy was to write for a more general audience, but to include a comprehensive set of references to literature at this interface. The statistical foundations of Gaussian process emulators are fairly standard, having been used in computer experiments across a wide range of subjects. With that in mind, we might add the following reference as a general and instructional introduction to the subject, for non-experts:

  O'Hagan, A. Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering & System Safety 91, no. 10 (2006): 1290-1300.

  Using emulators for climate science work is rarer, although a literature is building. Our paper uses standard emulators in a less standard way, in order to learn about the model and the climate. We can expand the literature review to include more examples of emulators being used in novel ways in climate science, in order to include more context for the reader. Some specific examples of related analyses from the climate science literature are:

  Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L. and Yamazaki, K., 2013. History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. Climate dynamics, 41(7-8), pp.1703-1729.

  Bounceur, N., Crucifix, M. and Wilkinson, R.D., 2015. Global sensitivity analysis of the climate-vegetation system to astronomical forcing: an emulator-based approach. Earth System Dynamics, 6(1), p.205.

  Tran, G.T., Oliver, K.I., Toal, D.J., Holden, P.B. and Edwards, N.R., 2016. Building a traceable climate model hierarchy with multi-level emulators. Advances in Statistical Climatology, Meteorology and Oceanography, 2(1), p.17.

- Action taken: We have added a section in the supplementary material with a brief outline of the emulator, more detail on the statistical modelling choices, and deeper reference to the software description paper. The introduction has been updated and streamlined, with these references added.

R1: Below are some comments that the authors might like to consider for a revised manuscript:
Overall points

- The title is possibly too general. The emphasis is on DGVM modelling of forests, not general overall issues of structure.

- Response: While we take the reviewers point here, we feel that the techniques used in the paper are sufficiently generalisable to be of interest to the wider climate modelling community. A key theme of this paper is that it attempts to improve the DVGM within the context of an Earth system model, which has it's own biases in climate simulation. An alternative title could be "The impact of structural error on parameter constraint in the land surface component of a climate model", but we welcome suggestions from the Editor.

- Action taken: None, as the Editor was happy with the original title.

- R1: The Abstract needs to be something that can be read in isolation, such that the reader can obtain a strong idea what the paper is about. To my mind, there is some repetition (e.g. three times says this uses "a history matching approach", and yet doesn't define what this actually is). Removing repetition can make space for more details. Extra description of the parameters changed would be helpful, rather than a vague "parameters that lead to a realistic forest fraction".

- Response: The reviewer makes some good points here, however describing the individual parameters in the abstract and yet making it shorter might be challenging. The focus of the paper is on the techniques for learning about the parameters, rather than the parameters themselves. Perhaps a broad description of the types of systems the parameters help control might be appropriate? We agree that the abstract could be more compact, avoid repetition and perhaps offer a clearer description of history matching. With that in mind, we suggest the following as a re-write:

  We use observations of forest fraction to constrain carbon cycle and land surface input parameters of the reduced resolution global climate model, FAMOUS. We use an ensemble of climate model runs to build a computationally cheap statistical proxy (emulator) of the climate model. We then use a "history matching" approach, comparing the emulated climate model output at various parameter settings, and ruling out as implausible those where the simulated output is judged statistically incompatible with observations. 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 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 calibrate the model leads to a larger underestimate of the Amazon forest fraction. We argue that this finding indicates a structural model discrepancy. We characterise this discrepancy, and explore the consequences of ignoring it in a history matching exercise. We use sensitivity analysis to find the parameters which have most impact

on simulator error. Finally, we perform a history matching exercise using credible estimates for simulator discrepancy and observational uncertainty terms. We are unable to constrain the parameters individually, but just under half of joint parameter space is ruled out 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.

- Action taken: Abstract re-written for clarity, slightly different from this version.

- R1: Reviewing this, I'm trying to really understand what the main thrust of this paper is about, in the statistical/algorithm sense. Can I confirm that the over-arching message is that quantity delta in Eqn (1) is important, can be characterised, and shows geographical variation. To my mind, that is a powerful result. It basically says if (i) not enough process representation is introduced in to a model, then structure deficiency gets masked by parameter fitting, and (ii) doing so will create problems between different locations.

- Response: That is a good summary of the main thrust of the paper. We would also like to highlight that it is not just missing process representation, but poor process representation (i.e. biases in other parts of the climate system) that can lead to errors if the delta in equation 1 is not taken into account. We would also like to highlight some of the novel techniques that we've developed to learn about discrepancy and its impacts. We will endeavour to make this clearer in the introduction to the paper.

- Action taken: Introduction edited for clarity.

- R1: It would be nice to acknowledge that structural errors presumably also reduce confidence in any model for future projections, even when just at a single region where it performs well for contemporary periods.

- Response: I suggest that we include this point in the discussion section. One advantage of including and estimating a discrepancy term is that future projections should acknowledge the uncertainty caused by the structural discrepancy. While this may lead to more uncertain projections, they should be more robust - that is, they should offer a more accurate estimate of uncertainty.

- Action taken: This point is made in the introduction, in the simulator discrepancy section.

- R1: Page 9, starting "Does this region represent". This is a critical part of the paper, discussing how in effect a standard best-fit might not always be appropriate. Can the discussion be led back to Eqn (1), and in particular the structural delta parameter? (Also line 1, page 9, I cannot see in a Table or diagram what the alternative potential values are, for comparison against the default inputs - apologies if I've missed something). Where are the local, or continent-scale, delta values given?

- Response: At the moment, this just says "without using a structural discrepancy function", but we agree with the reviewer that this could be much clearer. We will refer this straight back to equation 1, with the implications for mean and

uncertainty of the discrepancy function (not) used. The alternative potential values are a multidimensional cloud of points in parameter space, and therefore hard to summarise in a table (or even in a graphic - we are reduced to a two dimensional projection of the five dimensional space). The graphic (figure 4) has the space in normalised units - it might be clearer if we were to place the default parameters on this graphic. The model error in each forest at the input values indicated by figure 4 can be estimated by looking at figure 5, which shows the output of the model at this region.

- Action taken: Referred the section back to the equation 1. We have added the default parameter values to the graphic in figure 2, and to table 1 to make it easier for the reader to both find the values, and visualise them in comparison with the NROY regions in the other figures.

R1: Details

- P2, line 10. Again, please give the reader some idea what "History matching" is, given other quantities such as "calibration" and "tuning" are defined at this point.

- Response: We shall include an early, simple description of history matching, which may well be more unfamiliar than tuning or calibration to readers.

- Action taken: A clearer description of history matching has been added early in the introduction.

- R1: Around lines P2, lines 25-29. It would be really nice to have more concrete reasons why emulators, parameterisations etc are needed. This usually comes down to two factors: (1), computational speed prevents very high resolution modelling, even if the processes are more fully understood. For example, parameterisation of convection. (2), we don't know what the values should be, and these may exhibit strong regional heterogeneity. The latter is more the case for this paper, with questions asked as to what are the appropriate number of plant functional types that should be in land surface models - and if the number is high, can for example EO provide the values.

- Response: The reviewer makes a good point that we don't discuss the possibility of regionally varying parameters, and what that means for the current analysis. We shall include a section on regionally varying parameters in the discussion section, and expand the section on paramaterisation accordingly.

- Action taken: The suggested examples have been added to the introduction section.

- R1: Check notation is consistent throughout. P3, line 23, FAMOUS is described as a "climate simulator". In the minds of the authors, is this different to a standard GCMs. Do they regard FAMOUS's reduced resolution as removing it from being regarded as a full GCM?

- Response: In the statistical emulator literature "simulator" is often used for computational process models in order to distinguish them from statistical models. We will make this clear, and review use of "simulator" and "model" in the paper in order to ensure consistency.

– Action taken: Added footnote to this effect, and use of "simulator" is more standardised through the text now.

– R1: Again, in Section 1.3, this is now the 7th or 8th time that "history matching" is mentioned - it would be good to help the reader as to what it is, even if it is only to provide a methodological citation at this point.

– Response: See previous response.

– Action taken: Earlier history matching description added.

– R1: Cox (2001) is a technical note. Better to give a peer-reviewed reference?

– Response: It is possible to cite well known papers that use TRIFFID (e.g. Cox 2000), but Cox (2001) is the the standard reference that outlines the technical details of TRIFFID.

Cox, P.M., Betts, R.A., Jones, C.D., Spall, S.A. and Totterdell, I.J., 2000. Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. Nature, 408(6809), pp.184-187.

– Action taken: none.

– R1: P5, line 1. I don't understand the context of the sentence: "The Amazon region is not wet enough for a fully humid region to exist". If this refers to the FAMOUS model, and in particular its atmospheric response, then this will make any DGVM fail if rainfall totals are too small. P5, discussion of beta parameter. In a similar vain to the comment above, is it OK to treat the atmospheric beta parameter as a "nuisance" parameter? Isn't there a risk that errors in GCM-projected precipitation - for example - will affect best-fit parameters in Table 1?

– Response: FAMOUS has known biases, including a climatologically dry Amazon region, and this is indeed one of the strong candidates for low forest fractions in that region, as discussed later in the paper. However, in the Amazon region there are also possible confounding feedbacks between land cover and climate, making attribution of any biases more difficult. No climate simulation is perfect, and biases large or small are a common problem to be dealt with in any analysis. Our analysis offers new techniques to identify and characterise such biases, and the way that they might impact our estimates of the values of input parameters. The beta parameter is not correlated with any of the land surface parameters in the ensemble design, and so we felt justified in excluding it from analysis of the land surface parameters. However, it may well have an impact on climatology, and this could be the subject of a future study.

– Action taken: None.

– R1: P5, line 18. From code that is shared with other centres, TRIFFID has a rapid spin-up option to near-equilibrium. Does it really need 10000 years?

– Response: The fast spin-up mode was used in the simulations - only the equivalent of 10,000 years for each decade was used in this mode. The climate simulations were the averages of the last 30 years of a 200 year run. We shall make this clearer in the text.

- Action taken: text amended to reflect the reviewer's suggestion.

- R1: Trivial thing, but it might be nice in Figure 2 to write as S.E.Asia (not SEASIA).

- Response: This will be amended to be consistent with the other plots (a space added).

- Action taken: None, as the headings are directly taken from the R data frame that contains the data. Keeping a "no spaces" name ensures consistency with all of the other parameters in this diagram, and offers a direct check that we are plotting the correct thing.

- R1: Can I confirm that a reader could find all details of the emulator in the Roustant et al 2012 paper. So, for instance, what a "leave-one-out cross validation metric" is.

- Response: Roustant et al. (2012) is very comprehensive in its mathematical description of the emulator, and the software package that it informs. Leave-one-out cross validation is not related to the emulator itself, but is a broader validation algorithm. We will include a suitable reference (e.g. Hastie, Tibshirani and Friedman (2001).

  Hastie, T., Tibshirani, R. and Friedman, J. 2001. The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.

- Action taken: [as above] We have added a section in the supplementary material with a brief outline of the emulator, more detail on the statistical modelling choices, and deeper reference to the software description paper. We have added the latest version of the reference (Hastie et al 2009) to cross validation, and other statistical model verification techniques.

- R1: Figure 7 I find very useful as it allows assessment of the geographical differences, providing more information that the global parameterisation Table 3. There are quite a few statistical methods available to determine parameter importance and/or nuisance parameters. An extra sentence stating what additional benefit the FAST algorithm brings would be helpful - i.e. beyond just the Saltelli reference.

- Response: The FAST algorithm is ideally suited to our situation in that a) it provides an accurate global sensitivity analysis, including main effects and interaction terms and b) is easily and cheaply calculated using the emulator and a convenient R package. We shall include a sentence to this effect in the section.

- Action taken: Further justification for using the FAST algorithm added.

- R1: Figure 8 is important as it shows how the Amazon has a difference response. Or put another way, a calibration of NL0 and V_CRIT_ALPHA for the Amazon could find a pair of parameters that would clearly be sub-optimal when applied to the other 3 regions. And vice-versa. I'd like to see more discussion around Figure 8, how it demonstrates the structural problems (i.e. very different responses to NL0 and V_CRIT_ALPHA, depending on location), and again - can this be related back to the delta parameter? This will also link better to the paper title, which is about model structural problems.

- Response: Linking this clearly back to the structural discrepancy function at this point is a good idea. However, the discussion that the reviewer requests here is a large part of the later analysis (e.g. figures 10 - 12, and section 3.5). We could indicate the more detailed discussion in this later section in the text of this earlier section.

- Action taken: We've clarified the fact that non-overlapping regions of zero error in this reduced input space makes it likely we'll need a discrepancy function.

- R1: Figure 13 is nice and clear, and in many ways it is a shame that the paper is so long in technical details before getting to that point. Obviously this is a slightly naive comment, but could it simply be that the trees of the Amazon have evolved differently to those of Africa. This could possibly be due to different imposed climatologies that the trees have adapted/acclimated to. So one conclusion of this paper could simply be that any land surface model such as TRIFFID requires a parameter mask, or ancillary fields, that are different for different places. The paper hints at this, page 16, in "Causes of discrepancy", where different rooting depths are considered. One future work extension might therefore be to include a root depth as a geographically-varying parameter, to add to those in Table 1? Would this then collapse delta down to zero for all locations?

- Response: This interesting and useful idea should clearly be included in the discussion section.

- Action taken: This suggestion is included in the discussion, under the "Causes of discrepancy" section.

Reviewer 2 (Richard Wilkinson, RW)

- RW: This paper describes a thorough and detailed investigation into the ability of FAMOUS to predict forest fraction. The paper starts from the pretext of being given an ensemble of pre-run simulator evaluations and observation data corresponding to some of the outputs, and being asked to estimate some of the parameters. The work applies the latest statistical thinking/methodology in a largely clear and careful manner. To my non-climate trained eye, the authors seem to learn things about FAMOUS that were possibly unknown before, and likely to be of interest to the community of climate modellers. In my opinion the work deserves to be published subject to a few minor changes.

  I have two main criticisms of the paper. The first is that it is slightly repetitive in places. Several of the plots show very similar information, and make the same point albeit in different ways (which may be the intention). I felt the main point of the paper could be made in less space, and that this would improve the paper. My second criticism is that the paper is philosophically confused in places. This isn't necessarily a criticism of the paper, as most of the computer experiments community is somewhat confused about model discrepancy (as am I), but I felt the discussion lacked depth and nuance in places. Note that many of the following points are discussion rather than suggested changes to the manuscript.

- Response: Richard makes some valid points here, but the paper is long because it shows a number of novel analysis techniques, each of which provide some unique information about the simulator, its errors, and the relationship between the input parameters and the simulator output. Finding a clear narrative that included these analyses was a challenge,

but valuable. Excluding some of these analyses may well make the paper clearer in its main message, but at the risk of changing the focus on explanatory analyses, which I feel is a strength of the paper. However, I think it would be possible to move some of the analyses to supplementary material, if that was deemed beneficial. Sections such as 3.2 (sensitivity analysis) and 3.4 (How much space is ruled out by combinations of observations?), and parts of section 3.5 (e.g. figure 11), are somewhat additional to the main arguments of the paper, and could be moved. Both reviewers have made suggestions for expanding the discussion, which it is hoped will add depth and nuance.

- Actions taken: Removed discussion of sum of sensitivity effects (and related table 4), as it is a distraction. R1 says that more specific sensitivity is useful, so that section remains. Each section has been reviewed, to tighten language and to purge repetition. Some plots and analysis have been removed, and some moved to supplementary material, so the main arguments of the paper are now delivered in a shorter paper.

- RW: Simulator discrepancy. As discussed, estimating simulator discrepancy is hard, as it is difficult to disentangle the effect of simulator discrepancy from the problem of estimating unknown parameters. I dont like the definition of discrepancy quoted from Williamson et al 2014, that discrepancy is an error that cannot be removed by changing the parameters without introducing more serious biases to the model. The problem is that what constitutes an acceptable discrepancy function depends upon your goal. If you aim to do prediction, then something like the above would work, as we just want to characterize the simulator error for a given parameter value. However, if the aim is to infer the parameters, and for that inference to relate to the "true" value of those parameters, then you have to aim to model the true simulator discrepancy, which is much much harder. The problem that is hard to overcome, is that we may find the smallest simulator error occurs at parameters that are far from their "true' values if the simulator is poor. Brynjarsdottir and O'Hagan make the point that strong prior information is needed on the true parameter values if you wish to have any hope of disentangling the parametric uncertainty from the discrepancy. I think the aim of this paper is to estimate parameters, but the approach taken is one that is perhaps better suited to prediction problems.

  A discrepancy emerges in the paper, and is argued for by showing that there is an irresolvable error. The argument used is a kind of minimum error argument: we can't simulate all four forests simultaneously, but we can do three, so let's have a discrepancy just on the Amazon, and assume the simulator is fine for the others. This sounds sensible, but it could be that the Amazon is correct and the others wrong, or that there is simulator discrepancy for all four when we use the true parameter values. I could imagine that the errors are highly correlated for the forests, so that this kind of weight of evidence approach may be flawed. This also highlights for me the weakness of this approach compared to a more traditional statistical approach. If we had statistically modelled the discrepancy, described priors, and inferred posteriors, I suspect a similar conclusion may have been reached, but the weighting would have been done using the rules of probability, and the argument would instead be over the choice of model. Here, although it is unclear to me quite how the conclusion was reached, it seems that the authors avoid the need for modelling assumptions, but instead use an informal and heuristic weighting arguments to decide where to place the discrepancy. Although they have a mechanistic explanation of why their approach makes sense, the danger is that this is done post-hoc to fit the results.

A final point on the discrepancy concerns the sentence "We do not have enough information to create a more detailed discrepancy function: for example, one that varies across parameter space". Why would the discrepancy vary across parameter space? I thought it was the difference between the simulator and reality when the simulator is run at the "true" or "best" input?

– Response: The aim of this analysis is indeed to find good parameter sets, but also to use information that comes from the analysis to characterise the simulator discrepancy, and its consequences. It should perhaps be seen as a valuable, and useful step along the road towards a full statistical calibration treatment of the problem, rather than an end point. History matching is conceptually simple, easy to code, and fast to calculate, making it an accessible and attractive option for introducing more statistically robust analyses. With this in mind, we could expand the discussion to include some of the points made by Richard here - particularly the advantages of using a full calibration, compared to our more ad hoc approach.

  There is also no doubt that, even without a full calibration exercise, there is information that can be used to make judgements within the history matching framework. If we know that the model contains climate biases, and that parameters are the result of a long modelling effort and knowledge, do we rule them out as implausible when the model does not reproduce the Amazon? Very likely we would not, especially when other forests are adequately modelled.

Regarding the discrepancy varying across parameter space - this is incorrect, and we should remove the statement.

– Actions taken: Statement about variation across parameter space removed from introduction. Added an alternative definition of discrepancy under the best input approach, from e.g. Goldstein et al (2009) to introduction. Added the point that strong priors are needed to distinguish parameter uncertainty from discrepancy. Included a section in the discussion suggesting that a full probabilistic calibration could have advantages over our approach, particularly with regard to weighting inputs.

– RW: History matching. In the statistical part of the computer experiment community, there is an ongoing debate about whether we should do calibration or history matching (HM). I sometimes feel that HM advocates are too critical of calibration, criticising implementation problems as if they were fundamental flaws in the framework, and conversely that the calibration crowd simply don't consider doing anything different. I like the idea of history matching, and have used it in my own work, but my understanding is that it was developed for situations where you have a huge input space, most of which is implausible, which can then mean that it is hard to accurately emulate the simulator across the entire input space. If this is the case, conservatively ruling out parts of space in a sequence of HM waves, can make emulation much easier. I have heard HM advocates then say that they might finish the analysis with a calibration, which again makes sense to me, as this can provide more nuanced information along the lines of "we can't rule out $\theta = 2$, but it is much less likely than $\theta = 3$", which are statements that cannot be made within a HM approach. For the situation considered in this paper, there is no need to do waves of HM, as the emulator is adequate, and the data are such that only a small proportion of space can be ruled out (43% ruled out in the end). I can't help but feel that statistical calibration would

have been the better approach in this case (although this is a matter of taste). Indeed, although the authors provides a brief explanation of why they prefer HM, in several places, the authors treat the output of their inference as if it were the result of a probabilistic calibration.

- Response: Again, we feel that history matching has something to offer as an accessible alternative to, and preparation for, full calibration. It is also a good platform for some of the "what if" type exploratory analyses that we conduct in the study.

- Action taken: See below.

- RW: For example, Figure 16 is misleading. The histogram is suggestive of this being a distribution over the parameters. But as history matching was used, not calibration, there is no relevant information about the relative weighting of the parameters. This error is compounded in the sentence "The relative frequency of NROY points is higher in some locations than others [...] suggesting a higher probability that the best estimates of the parameters is in these regions". No statement can be made about probability here, as no probabilities were used and so this is misleading.

- Response: In hindsight, Richard is right here, and we should make more effort to make sure that our analyses are not interpreted as fully probabilistic. This will include removing figure 16 and associated text, and clarifying figure captions containing histograms.

- Action taken: Clarified the description of history matching, and its uses, to distinguish it better from calibration. Removed Figure 16 and associated text. Added a section in the discussion, pointing out that calibration would offer tools to move beyond the challenges described in this paper.

- RW: Line 6-8 on page 9 puzzled me, and also made me think that probabilistic calibration was perhaps what the authors had in mind. The claim is that finding the NROY region is near the edge of parameter space suggests a discrepancy function. I didn't really understand why this should be so, unless there is a secret/undeclared prior distribution that the authors have in mind, and that they believe the parameters really lie near the middle of the a priori plausible region. Of course, in a HM approach these consideration are not taken into account.

- Response: This argument may be the result of two things: 1) the structure of the study, where a pre-computed ensemble has been passed along to a (mostly) new set of authors, with little or no opportunity to re-run. In this case, the ensemble range is being used as the de facto plausible range of parameter values, which is perhaps not what was intended. This might mean that plausible parameter settings are to be found outside of the initial parameter space, and that plausible parameters are found against the very edge of parameter space. 2) As in this case, there is a suspicion that there is a discrepancy (a low Amazon forest fraction, perhaps caused by a climate bias), and some default parameters near the centre of the space, but no firm evidence until the analysis is run. The modellers then have to make a judgement as to whether applying a discrepancy term, or excluding the default parameters is appropriate, which is explored in the discussion section.

- Action take: Problematic sentences removed.

- RW: On page 7, line 10, the authors say that the "key" difference between calibration and HM is that points are not-ruled-out-yet (NROY) rather than "accepted". I find this point to be rather pedantic, as it is just a matter of labelling. I would say the key difference is that HM classifies points, but calibration describes a probability distribution over them. If we did calibration with uniform priors and thresholded the likelihood (using a pseudo- likelihood of either 0 or 1), then the two approaches can be made algorithmically equivalent (the interpretation remains different).

- Response: This point is well made, and we shall amend the text.

- Action taken: Text amended to clarify differences between history matching and calibration.

- RW: Finally, HM uses the implausibility given by equation 2 to score points, and then rejects points with a high score. We know from the theory of scoring rules that it is important to use a proper score, yet we can show that this score is improper (e.g. Gneiting and Raftery, JASA, 2007). Why doesn't this matter? We could use other scores in HM, and cut-offs other than the 3 sigma rule, and indeed on page 11, line 26-30, variations on how to threshold the plausibility are discussed. I support the authors' call for more research on the behaviour of the measures of implausibility, and perhaps suggest that links to scoring rules are investigated.

- Action taken - none

Other points

- RW: Page 8, line 27. Where does the 0.05 observation error come from? And the sentence "This corresponds to an expectation that the true 95% CI of ± 0.15" is incorrect I think. Pukelsheim's rule says that the 95% CI is contained within ± 0.15, not that it is equal to it. For a Gaussian rv, this would be a 99% CI for example.

- Response: The 0.05 observation error is an expert judgement of the true observational uncertainty, and a useful illustrative value, given that there is little information on the uncertainty of the observations themselves. We will make this clear in the text.

- Action taken: Now reads "On the advice of domain experts, we assume observational uncertainty of 0.05 (one standard deviation) 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 we contend that it would be hard to argue that this represents an over-constraint."

- RW: There is some confusion over the projections of points in the plots. In figure 8 for example, error is shown as a function of two parameters, where the effect of the other parameters has been averaged out. Is this useful? Just because the average error is zero, doesn't mean the error is zero anywhere. I appreciate this probably isn't what is happening, but the plots aren't necessarily a good idea.

- Response: The scatter (in colour) of the plots gives a visual impression of how much the error varies across the other parameters, and we would argue that the plots give a good indication of the regions of likely small error, even if they do not show where (or if) the error is exactly zero.

- Action taken: None

- RW: Page 11, line 14. I don't understand the final sentence here? According to equation 2, it makes no difference whether we assign the uncertainty to the observation or the model discrepancy. And why would we want to do this? We were told observation error was known (and fixed).

- Response: We should make it clearer that the observational error is assumed, an expert judgement, and that arguments could be made for other values.

- Action taken: text amended to clarify that the observational and structural discrepancy uncertainty are assumed in this part of the experiment, but that the emulator uncertainty is emergent.

- RW: Another point that is more discussion then criticism, as I believe it is probably common practice, is the issue of treating the climate as a static system, by spinning up the climate model to reach equilibrium. Again, I'm not a climate scientist, but as the climate is dynamic, does this practice cause a bias? Suppose we had the true simulator, with zero discrepancy, would spinning-up to equilibrium induce an error in our predictions? I appreciate there is probably no way around this.

- Response: The practice of spinning up a model is a useful way to remove biases, given that we very likely do not have adequate data to initialise the entire state of the climate system. A spin up to some historical state, followed by a period of historical forcing can be used to get the state and the dynamics of the system correct, before predictions are made.

- Action taken: none.

- RW: The language needs editing in places, with errors becoming increasingly common in later sections.

- Action taken: Language was reviewed, corrected and simplified throughout the manuscript.

Minor points

- RW: Page 1, line 10, "find the parameters that have most impact on simulator error". To be slightly nit-picky, I don't know what this means. Perhaps "find the parameters that have most impact on simulator output", as simulator error, probably means the error when run at the best input.

- Response: This was an attempt to be compact, but we will correct to make this clearer.

- Action taken: language clarified - now suggest that we find those parameters that have the largest impact on simulator output.

- RW: Page 2, line 8-11. This description is slightly confusing. Calibration, tuning, and history matching are all solving the inverse problem in some sense. Needs rephrasing, and perhaps a reference or two.
- Response: We will rephrase this section as suggested.
- Action taken: The section was rephrased for clarity.
- RW: Page 5, line 7-8. I don't believe the claim that LHC designs are better than others. I read Urban and Fricker a long time ago, but I think they just compared LHC to grid designs, and then only in empirical experiments. I'm pretty sure it is not the case that the question of the best design is settled in general (see Zhu and Stein 2006, and Zimmerman 2006 etc). I think it would be better to say that LHC designs are "good designs".
- Response: We will rephrase this section as suggested.
- Action taken: We have stated simply that these designs are commonly used to construct emulators.
- RW Page 6, line 25, "Gaussian' not "gaussian"
- RW Page 6, line 29, "The emulator is a nonlinear regression model' perhaps "non-parametric" would be better than "nonlinear", given the potential for confusion with what is normally meant by "nonlinear regression model" i.e., non-linear in the Parameters.
- Response: We will rephrase as suggested.
- Action taken: both rephrased.
- RW Page 6, line 31. Given it is quite a long paper, there are remarkably few details about the emulator, covariance function, mean function, estimation approach etc. The review guidelines ask me to check that the paper is reproduceable, which without these details, it would not be.
- Response: These details could go in the supplementary material, along with the emulator verification.
- Action taken: These details are now in the supplementary material.
- RW: Page 8, line 31, "We sample from the emulator uniformly across input parameter space" - this is unclear. Presumably you sampled uniformly from the input parameter space, and then from the emulator. Same again on page 11, line 2.
- Action taken: rephrased for clarity
- RW: Page 10, line 7, "total effect"
- RW: Page 12, line 29. "that that model discrepancy uncertainty is zero".
- RW: Page 16, line 29/30. Rephrase sentence "First, is there..." A dodgy emulator would lead us to think a bias exists, not cause it.

- Response: We will rephrase as suggested.
- Action taken: All rephrased as suggested.

Author contributions. DM and all authors designed the analysis. DM conducted the analysis and wrote the paper. JW provided the FAMOUS ensemble and BB provided the observed forest fraction data.

Acknowledgements. This work was supported by the Joint UK ~~DECC~~BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101). DM was supported on secondment to Exeter University by the Met Office Academic Partnership (MOAP) for part of the work. JW was supported by funding from Statoil ASA, Norway. RB is a member of the editorial board of Earth System Dynamics.

References

Abramowitz, G.: Towards a public, standardized, diagnostic benchmarking system for land surface models, Geoscientific Model Development, 5, 819–827, doi:10.5194/gmd-5-819-2012, http://www.geosci-model-dev.net/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, Environmental Research Letters, 7, 024 002, http://stacks.iop.org/1748-9326/7/i=2/a=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 System Dynamics, 4, 95–108, doi:10.5194/esd-4-95-2013, http://www.earth-syst-dynam.net/4/95/2013/, 2013.

Bounceur, N., Crucifix, M., and Wilkinson, R.: Global sensitivity analysis of the climate-vegetation system to astronomical forcing: an emulator-based approach, Earth System Dynamics, 6, 205, http://www.earth-syst-dynam.net/6/205/2015/, 2015.

Brynjarsdóttir, J. and O'Hagan, A.: Learning about physical parameters: the importance of model discrepancy, Inverse Problems, 30, 114 007, http://stacks.iop.org/0266-5611/30/i=11/a=114007, 2014.

Carslaw, K., Lee, L., Reddington, C., Pringle, K., Rap, A., Forster, P., Mann, G., Spracklen, D., Woodhouse, M., Regayre, L., et al.: 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, Theoretical and Applied Climatology, 78, 137–156, doi:10.1007/s00704-004-0049-4, http://dx.doi.org/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, edited by Gatsonis, C., Hodges, J., Kass, R., McCulloch, R., Rossi, P., and Singpurwalla, N., vol. 3, pp. 36–93, Springer-Verlag, New York, USA, 1997.

Gnanadesikan, A. and Stouffer, R. J.: Diagnosing atmosphere-ocean general circulation model errors relevant to the terrestrial biosphere using the Koppen climate classification, Geophysical Research Letters, 33, n/a–n/a, doi:10.1029/2006GL028098, http://dx.doi.org/10.1029/2006GL028098, l22701, 2006.

Goldstein, M. and Rougier, J.: Reified Bayesian modelling and inference for physical systems, Journal of Statistical Planning and Inference, 139, 1221–1239, http://www.sciencedirect.com/science/article/pii/S0378375808003303, 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, Philosophical Transactions of the Royal Society of London B: Biological Sciences, 363, 1761–1766, doi:10.1098/rstb.2007.0024, http://rstb.royalsocietypublishing.org/content/363/1498/1761, 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, Climate Dynamics, 16, 147–168, doi:10.1007/s003820050010, http://dx.doi.org/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, Climate Dynamics, 37, 705–719, doi:10.1007/s00382-010-0934-8, http://dx.doi.org/10.1007/s00382-010-0934-8, 2010.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M.: Computer Model Calibration Using High-Dimensional Output, Journal of the American Statistical Association, 103, 570–583, doi:10.1198/016214507000000888, http://dx.doi.org/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 Dyn, 35, 785–806, doi:10.1007/s00382-009-0630-8, http://dx.doi.org/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, http://dx.doi.org/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, Climate Dynamics, 41, 2921–2936, doi:10.1007/s00382-012-1644-1, http://dx.doi.org/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, Climate Dynamics, 25, 189–204, doi:10.1007/s00382-005-0027-2, http://dx.doi.org/10.1007/s00382-005-0027-2, 2005.

Kennedy, M. and O'Hagan, A.: Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 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, Proceedings of the National Academy of Sciences, doi:10.1073/pnas.1507050113, http://www.pnas.org/content/early/2016/02/04/1507050113.abstract, 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, International Journal of Remote Sensing, 21, 1303–1330, doi:10.1080/014311600210191, http://dx.doi.org/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, http://www.biogeosciences.net/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, Proceedings of the National Academy of Sciences, 106, 20 610–20 615, doi:10.1073/pnas.0804619106, http://www.pnas.org/content/106/49/20610.abstract, 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, pp. 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, Geoscientific Model Development, 6, 1715–1728, doi:10.5194/gmd-6-1715-2013, http://www.geosci-model-dev.net/6/1715/2013/, 2013.

O'Hagan, A.: Bayesian analysis of computer code outputs: a tutorial, Reliability Engineering & System Safety, 91, 1290–1300, http://www.sciencedirect.com/science/article/pii/S0951832005002383, 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, Climate Dynamics, 16, 123–146, doi:10.1007/s003820050009, http://dx.doi.org/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, https://CRAN.R-project.org/package=sensitivity, r package version 1.11.1, 2015.

Pukelsheim, F.: The three sigma rule, The American Statistician, 48, 88–91, 1994.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2016.

Roustant, O., Ginsbourger, D., and Deville, Y.: DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization, Journal of Statistical Software, 51, 1–55, doi:10.18637/jss.v051.i01, https://www.jstatsoft.org/index.php/jss/article/view/v051i01, 2012.

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, http://amstat.tandfonline.com/doi/abs/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, Journal of Climate, 9, 676–705, doi:10.1175/1520-0442(1996)009<0676:ARLSPF>2.0.CO;2, http://dx.doi.org/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, Climate Dynamics, 38, 2513–2542, doi:10.1007/s00382-011-1208-9, http://dx.doi.org/10.1007/s00382-011-1208-9, 2011.

Smith, R. S.: The FAMOUS climate model (versions XFXWB and XFHCC): description update to version XDBUA, Geoscientific Model Development, 5, 269–276, doi:10.5194/gmd-5-269-2012, http://www.geosci-model-dev.net/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, Geoscientific Model Development, 1, 53–68, doi:10.5194/gmd-1-53-2008, http://www.geosci-model-dev.net/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, http://science.sciencemag.org/content/334/6053/230, 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, Advances in Statistical Climatology, Meteorology and Oceanography, 2, 17, http://www.adv-stat-clim-meteorol-oceanogr.net/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, Computers & Geosciences, 36, 746–755, 2010.

Vernon, I., Goldstein, M., and Bower, R.: Galaxy formation: a Bayesian uncertainty analysis, Bayesian Analysis, 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, http://dx.doi.org/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, Geoscientific Model Development, 6, 141–160, doi:10.5194/gmd-6-141-2013, http://www.geosci-model-dev.net/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, Geoscientific Model Development, 7, 1419–1431, doi:10.5194/gmd-7-1419-2014, http://www.geosci-model-dev.net/7/1419/2014/, 2014.

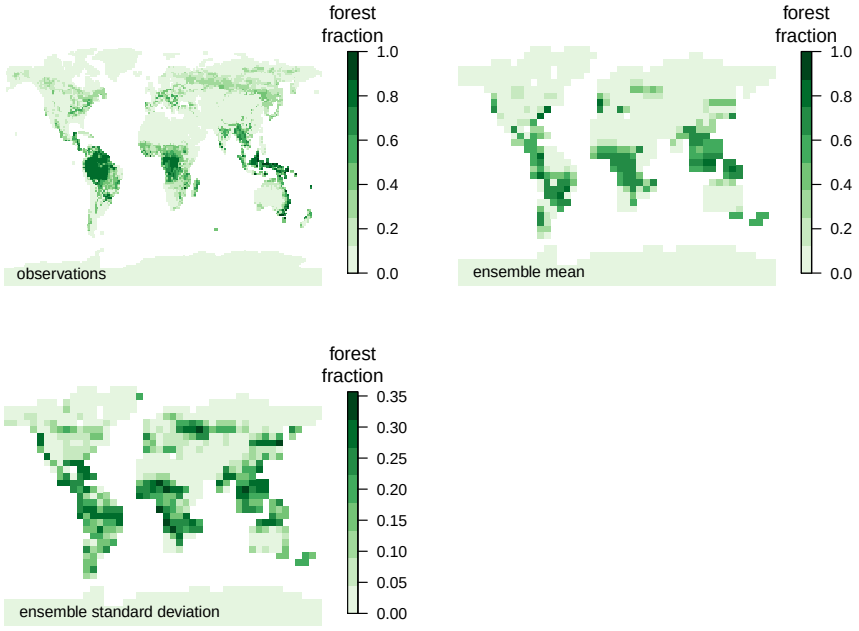

Figure 1. Observations of Broadleaf forest fraction (top left). Mean (top right) and standard deviation (bottom left) of broadleaf forest fraction across the 100 member ensemble of FAMOUS.

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, Climate dynamics, 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, Climate Dynamics, 45, 1299–1324, doi:10.1007/s00382-014-2378-z, http://dx.doi.org/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?, Climate Dynamics, 41, 3127–3143, doi:10.1007/s00382-012-1582-y, http://dx.doi.org/10.1007/s00382-012-1582-y, 2012.

~~Marginal histograms of the relative frequency of NROY emulated input points in each dimension of parameter space, using all forest observations and a discrepancy function for the Amazon.~~

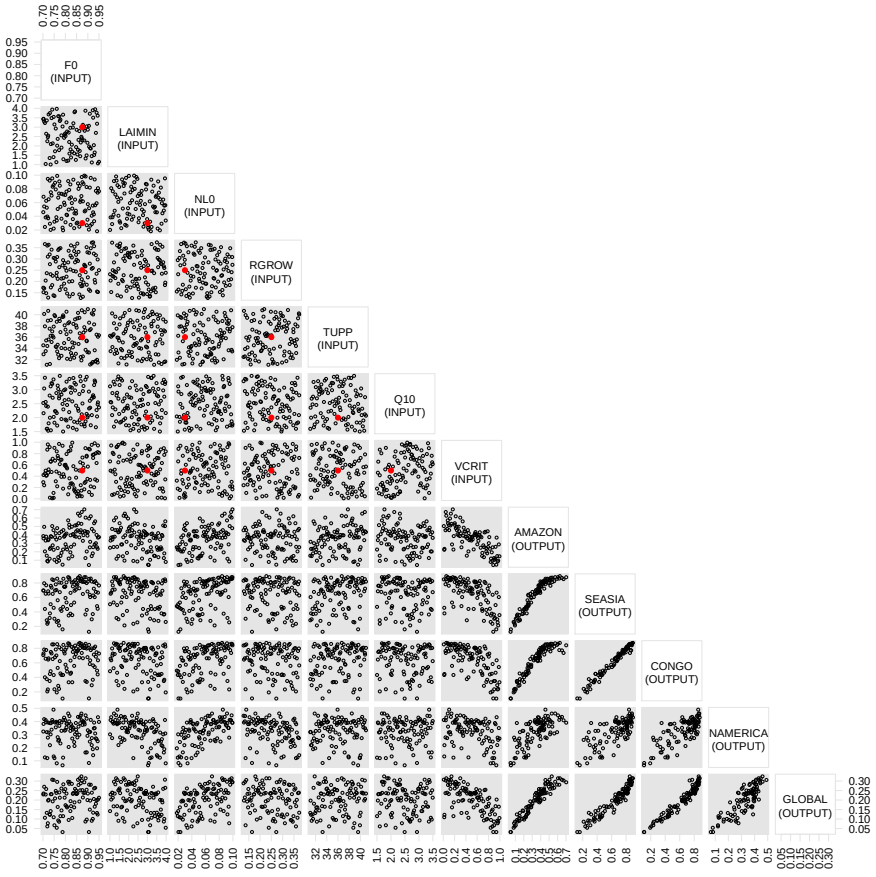

Figure 2. FAMOUS input parameters and forest fraction parameters, plotted against each other. Default inputs (not run) are marked in red.

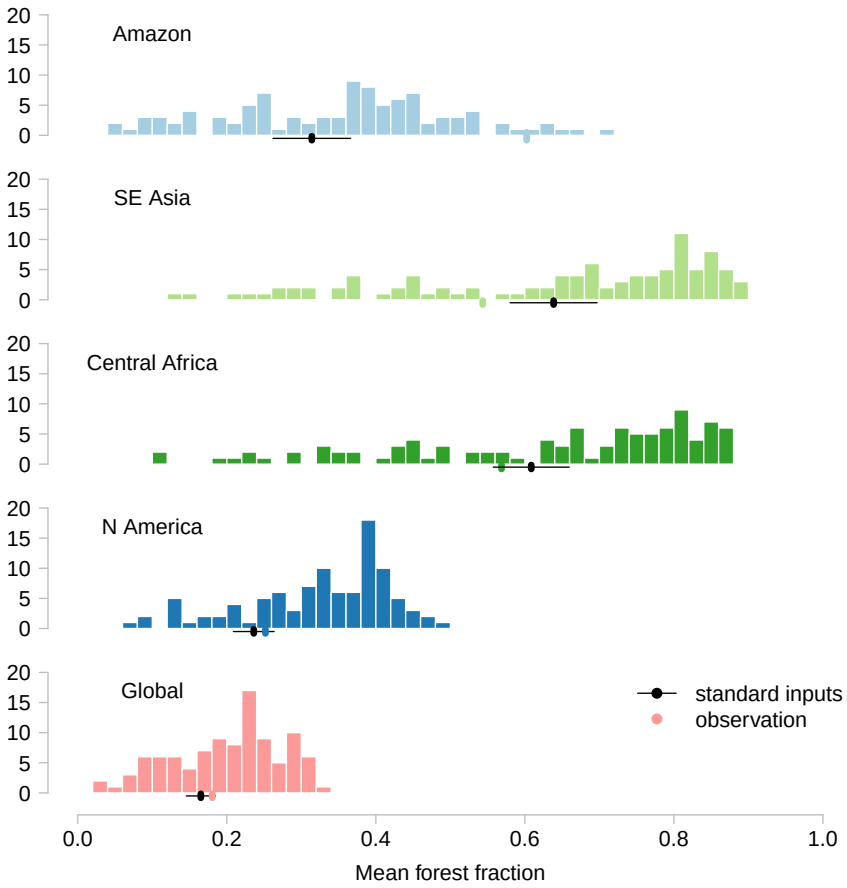

Figure 3. Histograms representing the number of ensemble members of a particular forest fraction in each region, and globally. Points plotted below the histograms represent the observed forest fraction (colours), ~~and12cm~~ and the forest fraction simulated at the "standard" parameters ± 1 standard deviation (black).

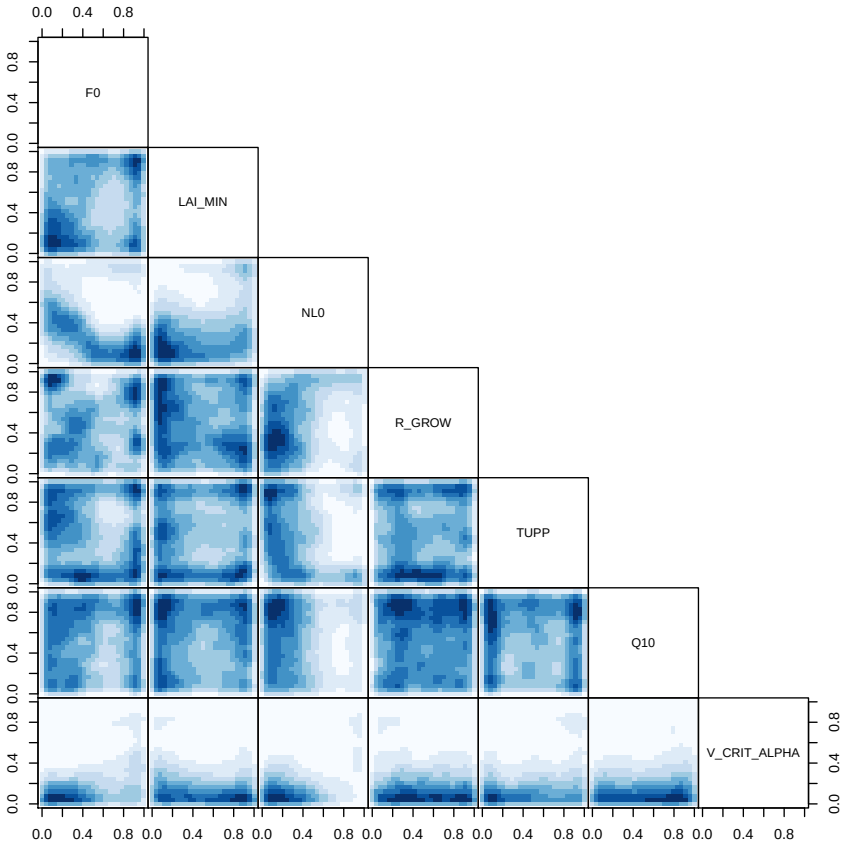

Figure 4. A density pairs plot of two dimensional projections of parameter space. The blue areas represent the density of NROY points, using all of the data, with an assumed observational uncertainty of 0.05 (1 standard deviation).

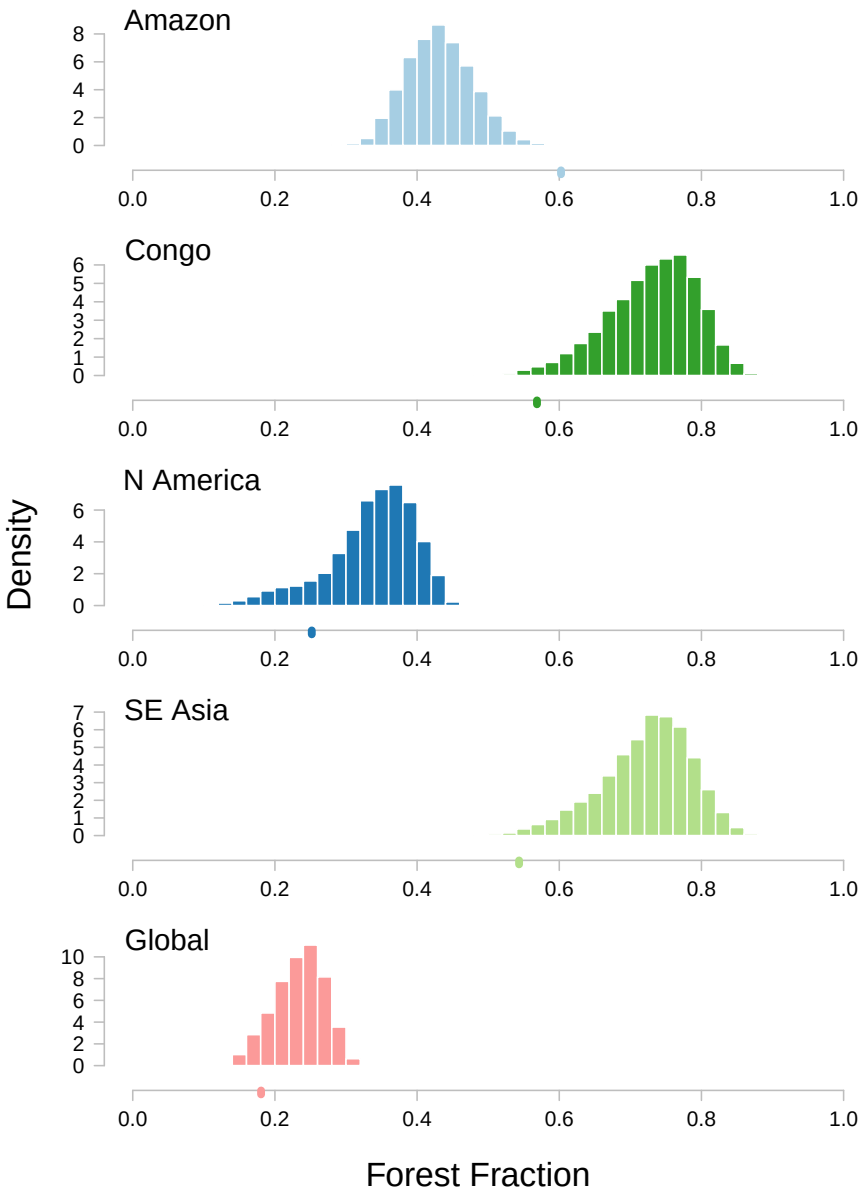

Figure 5. Best-estimate draws of forest fraction output from the emulator, at the set of points Not Ruled Out Yet when assuming a credible observational uncertainty. The value of the observed forest fractions are are plotted as a single point on the corresponding x-axes (a "rug plot").

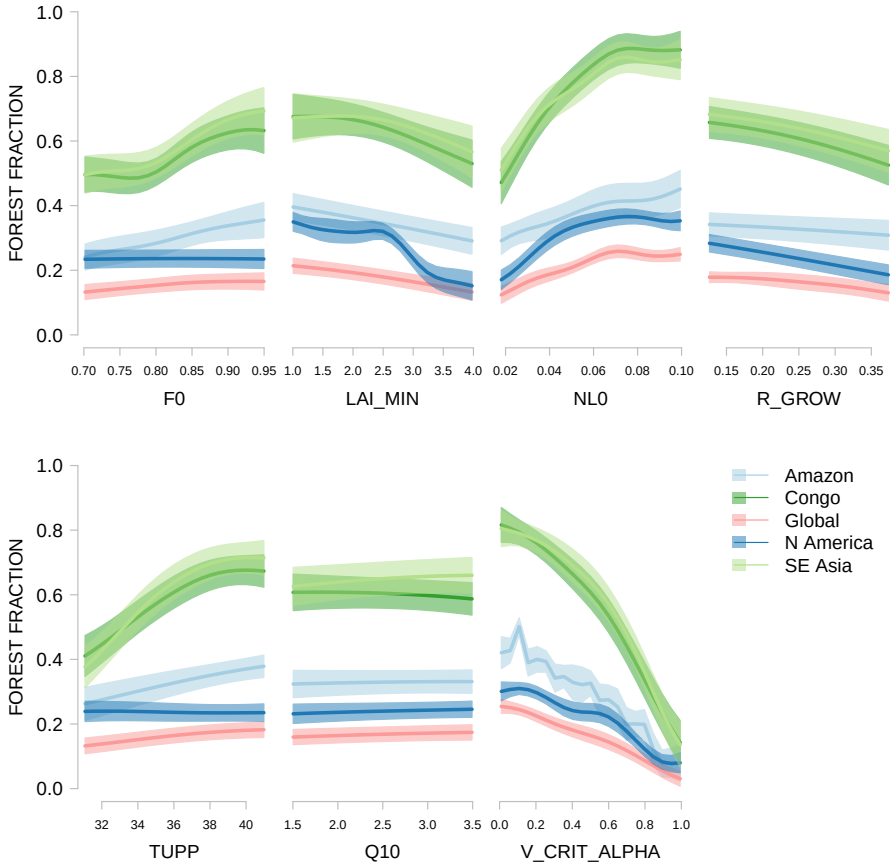

Figure 6. Marginal sensitivity of mean forest fraction to each input parameter in turn, with all other parameters held at standard values. Central lines represent the emulator mean, and shaded areas represent the estimate of emulator uncertainty, at the ± 1 standard deviation level.

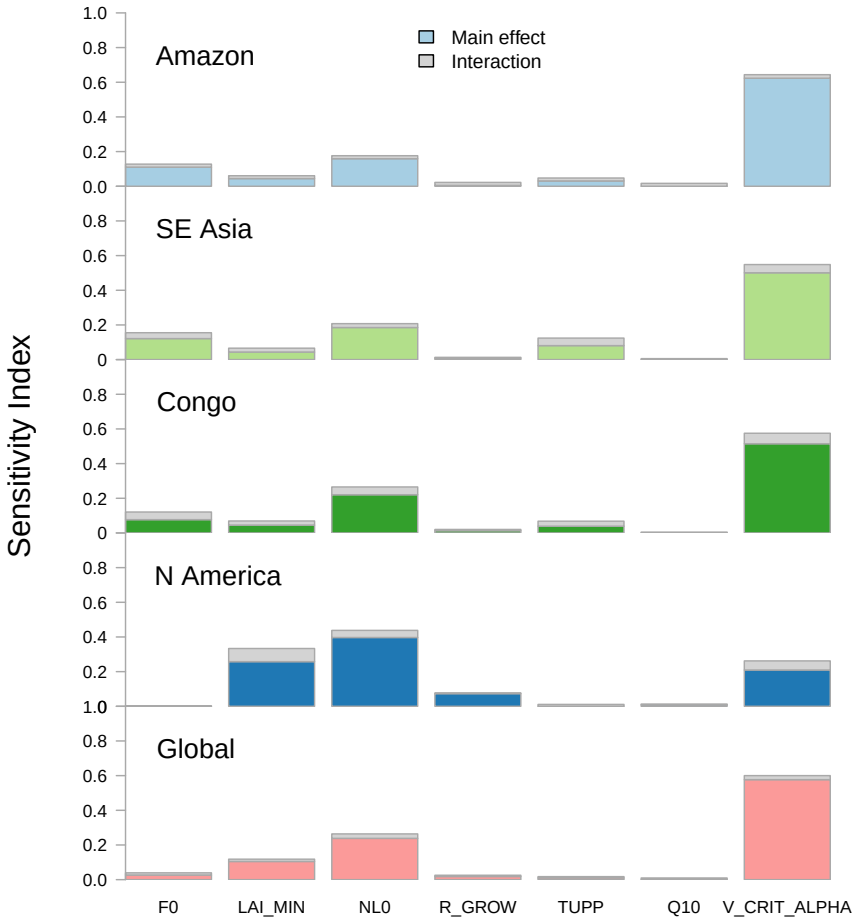

Figure 7. Sensitivity analysis of forest fraction via the FAST algorithm of **?**.

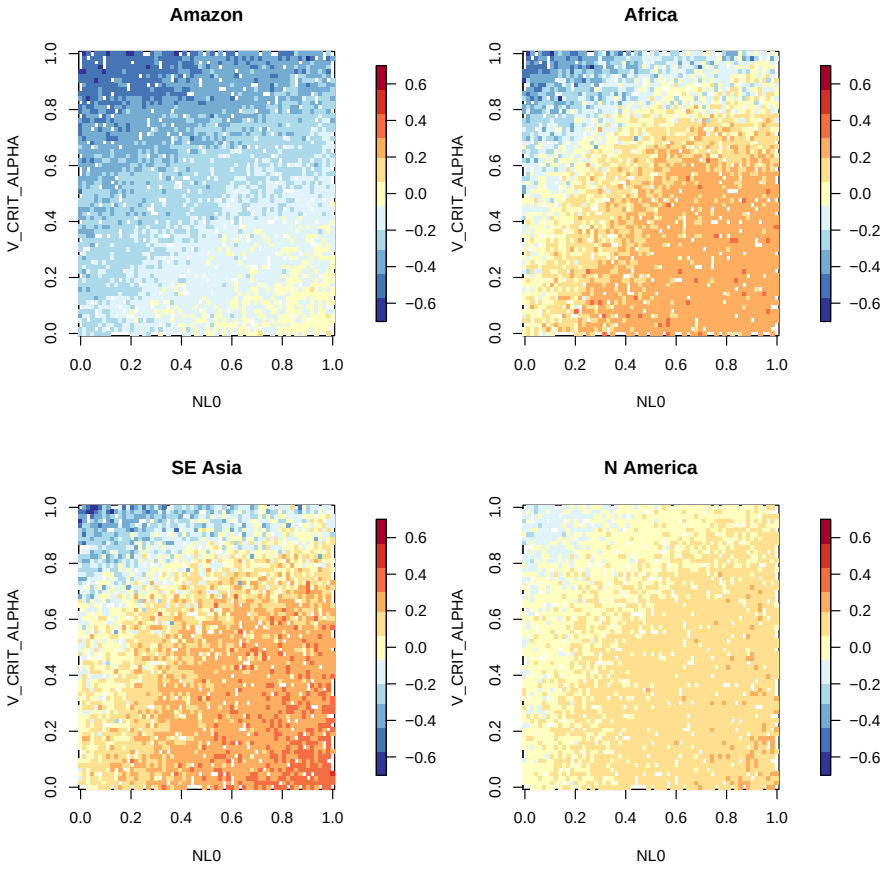

Figure 8. Maps of simulator error, in units of forest fraction, when projected into the two dimensional space of the most active parameters, NL0 and V_CRIT_ALPHA.

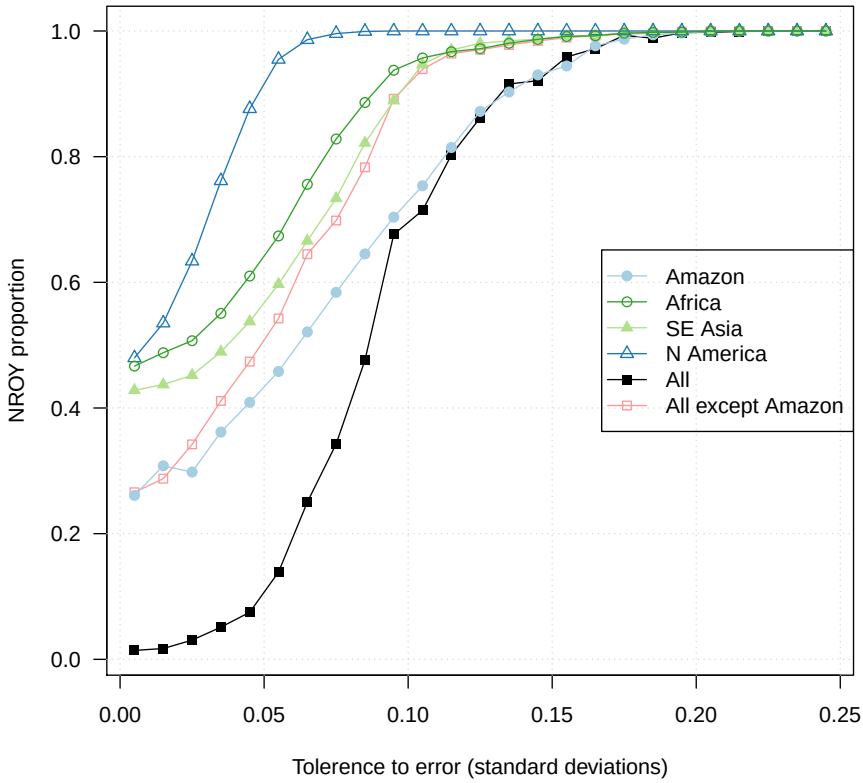

Figure 9. Proportion of NROY (Not Ruled Out Yet) input space plotted against "tolerence to error" - the total error budget including emulator, observational and simulator discrepancy uncertainty.

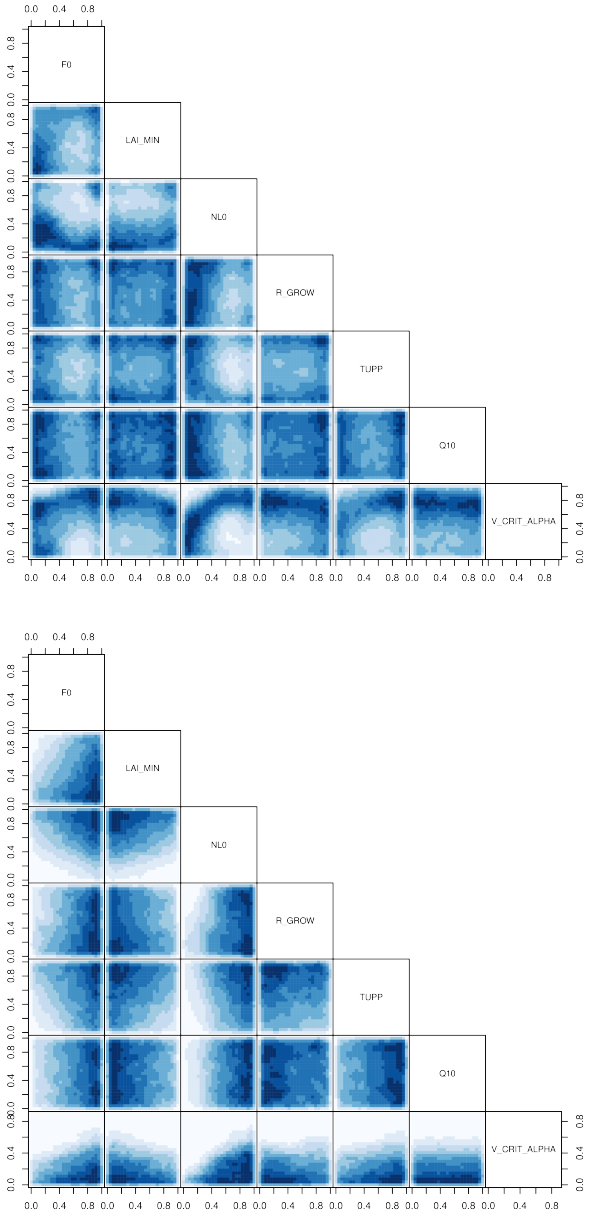

Figure 10. Marginal density of input parameter sets consistent with a very low "tolerance to error", and perfect observations, for the North American, Southeast Asian and Central African forests combined (top) and the Amazon (bottom). Dark blue regions indicate those with the highest concentration of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

~~Implausibility, given a "tolerance to error" of 0.1, varying two parameters at a time and holding all others at their default values. Amazon forest (top) and Central African forest (bottom). Blue indicates regions where the model best simulates the individual option, while red indicates regions where the model simulates the forests more poorly. The green point indicates the location of the "standard" set of parameters for FAMOUS in the varied dimensions.~~

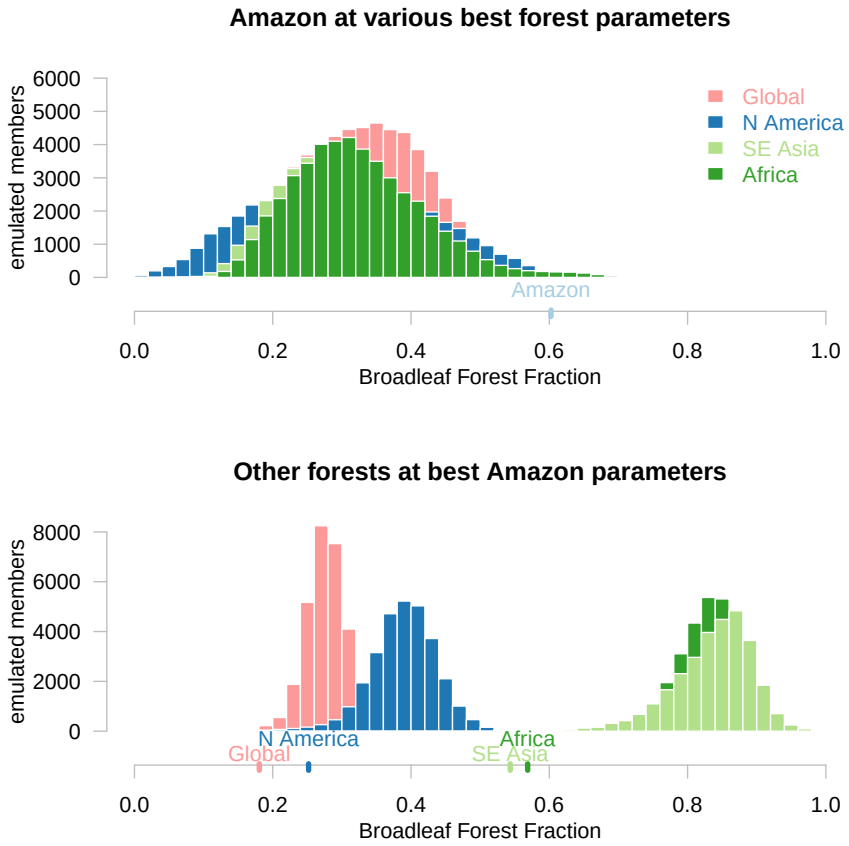

Figure 11. (Top) Forest fraction in the FAMOUS Amazon at the set of parameters where the FAMOUS best matches each of the other forest observations. (Bottom) Other forests in FAMOUS at the set where the FAMOUS Amazon best matches observations. Observed forest fractions are shown as marks underneath the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

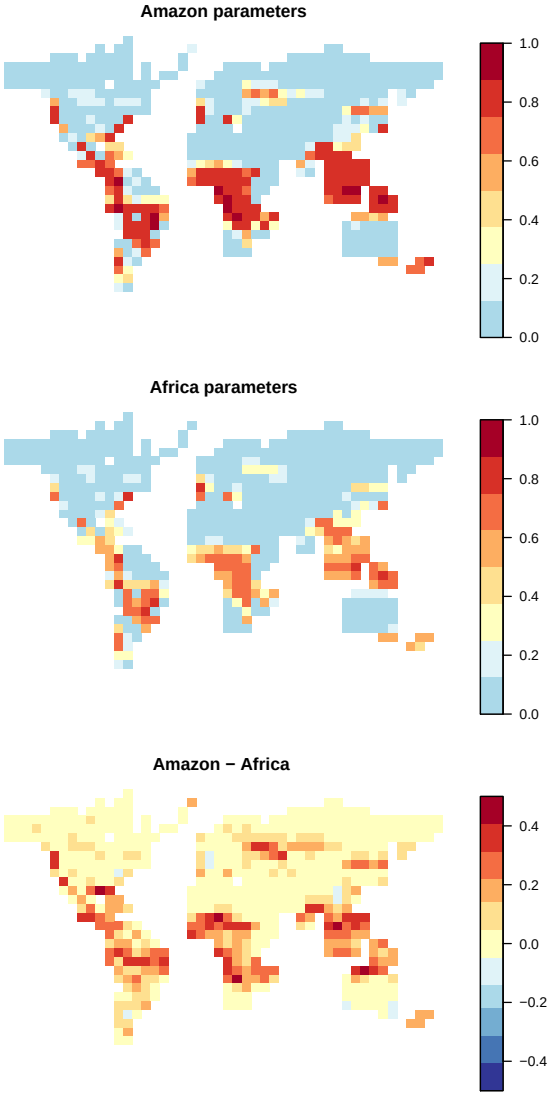

Figure 12. Maps of mean broadleaf forest fraction, over the "best" set of parameters found for the Amazon (top) and the Central African forest (centre). The difference between the two is mapped at the bottom.

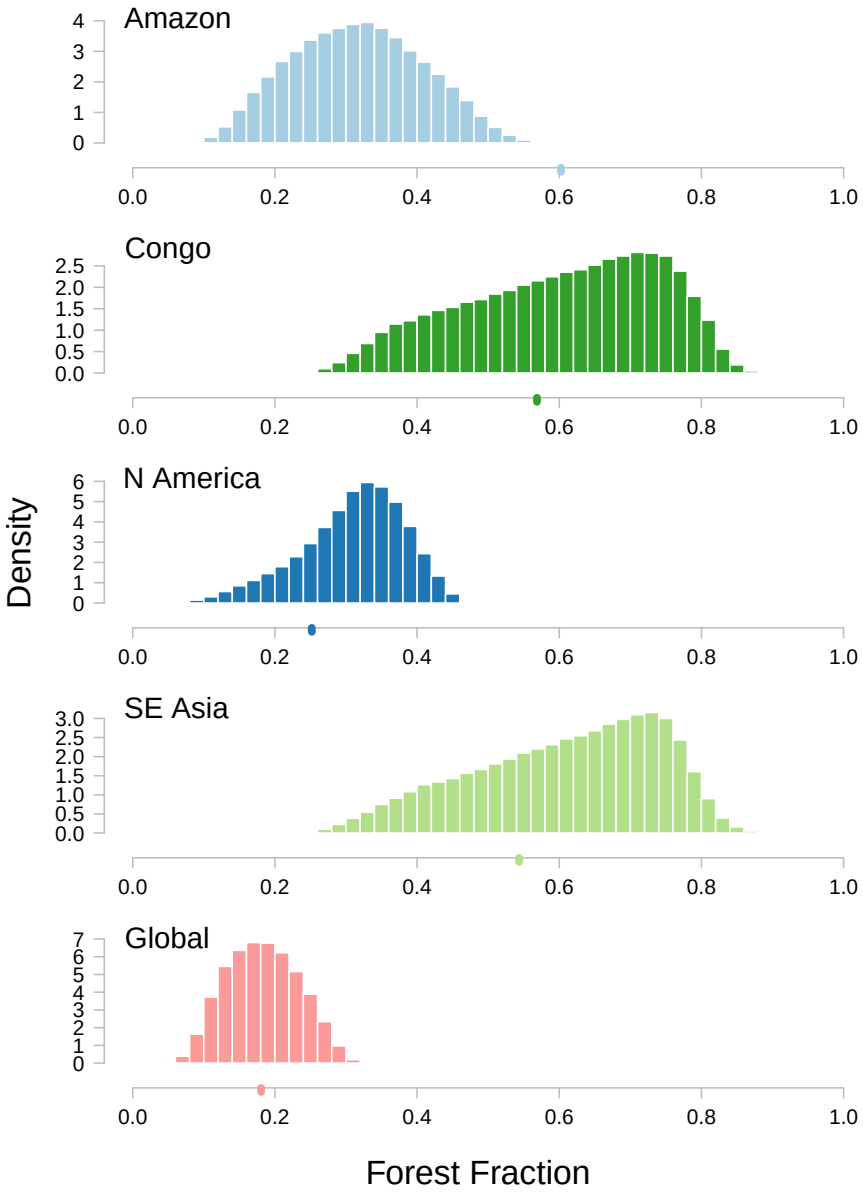

Figure 13. Histograms of emulated simulator output using credible estimates for observational uncertainty, a ~~model~~ simulator discrepancy term for the Amazon, and credible discrepancy uncertainty.

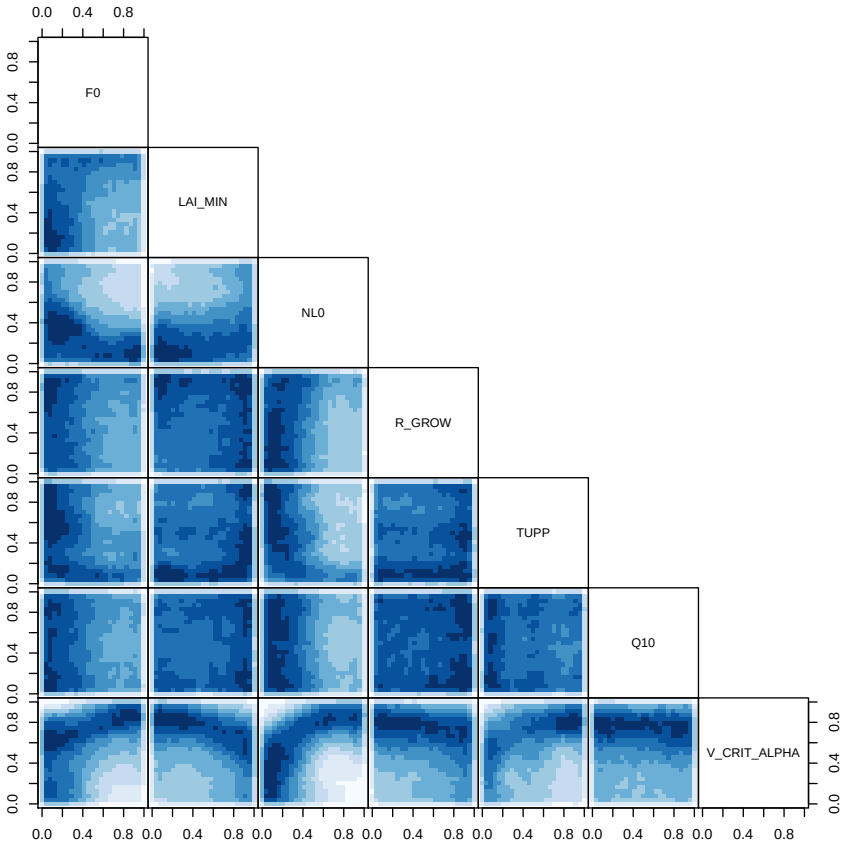

Figure 14. A density plot of the two dimensional projections of NROY samples from the design input space, using a all forest observations and a discrepancy function for the Amazon.