Correcting a bias in a climate model with an augmented emulator

Doug McNeall[1], Jonny Williams[2], Richard Betts[1,3], Ben Booth[1], and Peter Challenor[3]

[1]Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK
[2]NIWA
[3]University of Exeter

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

Abstract. We develop a method of bias correcting a climate model with a Gaussian process emulator, allowing valid values of input parameters to be found even in the presence of a significant model bias.

A previous study (McNeall et al. 2016) found that a climate model had to be run using land surface input parameter values from very different, almost non-overlapping parts of parameter space in order to satisfactorily simulate the Amazon and other forests respectively. As the forest fraction of the other forests in the model were broadly correct at the default parameter settings and the Amazon too low, that study suggested that the problem likely lay in the model's treatment of the Amazon region than the other regions. The study suggested that this might be due to (1) structural errors such as missing deep-rooting in the Amazon in the land surface component of the model, (2) a warm-dry bias in the Amazon climate of the model, or a combination of both.

In this study we bias correct the climate of the Amazon in a climate model using an "augmented" Gaussian process emulator, where variables often regarded as outputs of the model are treated as inputs. We treat the regional temperature and precipitation of the model as additional inputs to the emulator alongside the standard model inputs. We can then explore the relationship between climate, input parameters and the output of the emulator, forest fraction, finding that the forest fraction is nearly as sensitive to climate variables as any of the land surface inputs. Bias correcting the climate in the Amazon region using the emulator corrects the forest fraction to tolerable levels in the Amazon at many candidates for land surface input parameter values, including the default ones. It also increases the valid input space shared with that suggested by the other forests. We no longer need to invoke a structural model error in the Amazon, beyond having too dry and hot a climate.

Using the augmented emulator allows the bias correction a pre-existing coupled ensemble of climate model runs, reducing the risk of choosing poor parameter values because of an error in a sub-component of the model. We discuss the potential of the augmented emulator to act as a translational layer between model sub-components simplifying the process of model tuning when there are potential compensating errors, and helping model developers prioritise model errors to target. Our technique has the potential to help choose good input parameters for a model, and to efficiently project the impacts of a changing climate, even when there are significant biases in a sub-component of the model.

Copyright statement. TEXT

1 Introduction

1.1 Uncertainty quantification, compensating errors and model discrepancy

The field of Uncertainty Quantification (UQ) has seen a rapid development of methods to quantify uncertainties when using complex computer models to simulate real, physical systems. These models often contain simplifications of processes too complex to represent explicitly in the model, termed parameterisations. Associated with these parameterisations are coefficients called input parameters, the values of which are uncertain and can be set by the model developer. The settings have a material effect on the way the parameterisations operate, and therefore on output of the model, but often to an extent that is unknown until the model is run. Input parameters are subject to uncertainty and may be difficult or even impossible to observe, having no direct analogue in the real system. The process of setting the values of the input parameters so that the simulator output best matches the real system is called tuning, and where a probability distribution is assigned for the input parameters, it is termed calibration. Uncertainty in input parameters can induce uncertainty in the output of the model, leading to uncertainty in projections of future climate states or reconstructions of past ones.

Without strong prior information it can be difficult to attribute a difference between simulator output and the real system to underlying model errors, to an incorrect set of input parameters, or to inaccuracies in the observations. Similarly, there are often a number of ways to set parameters that lead to a particular model output, with poor choices of a particular parameter compensating for poor choices of other parameters, or for modelling errors. This situation means that a good candidate for input parameters might be found in a large volume of input space, and projections of the model made with candidates from across that space might display a very wide range of outcomes. This problem is sometimes referred to as “identifiability", but otherwise known as “equifinality", or the “degeneracy" of model error and parameter uncertainty. It can be relatively easy to find a good subset of input parameters given a small set of inputs and outputs and a well behaved relationship between the two. This situation might be found for a subcomponent of a climate model, where there are good observations of the system being studied, for example. Improving a coupled climate model however can require an involved and lengthy process of development. Some components of the model may have been tuned to compensate for errors in others or there may be unknown errors in the model or observations. Further, more complex models are computationally expensive and so infeasible to run in enough configurations to be able to identify these errors.

Hourdin et al. (2017) offer a summary of current practice in the somewhat understudied and sparsely documented field of climate model tuning. While there are clearly common features, there appear no standard procedures for climate model tuning however. As Hourdin et al point out, it remains an art as well as a science. Various individual centres have begun to document their tuning practices with regard to tuning targets and procedures (Schmidt et al. (2017), Zhao et al. (2018), Walters et al. (2017))

Parameter tuning occurs at different stages in model development, perhaps starting with single column version of the model. The climate model components to be coupled might be then tuned with standard boundary conditions - for example tuning a land/atmosphere component with fixed or historically observed sea surface temperatures. Finally, a system-wide tuning might be used to check that there are minimal problems once everything has been coupled together.

Golaz et al. (2013) Show the potential impact of compensating errors in tuning. They find that two different but plausible parameter configurations of the cloud formations of the coupled climate model GFDL-CM3 can result in similar present-day radiation balance. The configurations did not differ in their present day climate, but showed significantly different responses to historical forcing and therefore historical climate trajectories.

Although climate model tuning is overall a subjective process, individual parts are amenable to more algorithmic approaches. Statistical and machine learning approaches to choosing parameters to minimise modelling error, or to calculate probability distributions for parameters and model output are known as uncertainty quantification (UQ).

The problem of accounting for model discrepancy when using data to learn about input parameters is becoming more widely recognised in UQ. It was formalised in a Bayesian setting by Kennedy & O'Hagan (2001). The authors suggested simultaneously estimating a model discrepancy - there called model inadequacy - as a function of the inputs, using a Gaussian process prior.

Arndt et al. (2012a) offer a number of examples of identifiability problems, ranging from solvable using mild assumptions through to virtually impossible. In a companion paper (Arndt et al. 2012b), they outline a way of improving identifiability using multiple model responses.

Brynjarsdottir and O'Hagan (2016) argued that only by accounting for model discrepancy does even a very simple simulator have a chance of making accurate predictions. Further, they found that only where there is strong prior evidence about the nature of that model discrepancy is it possible to solve the inverse problem and recover the correct inputs. Without this strong prior evidence the estimate of the correct parameters is likely to be overconfident, and wrong, leading to overconfident and wrong predictions of out-of-sample data.

Some of the dangers of overconfident and wrong estimates of input parameters and model discrepancy can be reduced using a technique called history matching (Craig et al, 1996), sometimes called pre-calibration or iterated refocussing. The aim of history matching is not to find the most likely inputs, but to reject those unlikely to produce simulations statistically close to observations of the real system. An implausibility measure (I) is calculated, taking into account the distance between the simulator output and the observation, but allowing for uncertainty in the observations, the simulator output and the simulator discrepancy. Those inputs that produce a large implausibility score are ruled out from consideration as candidate points.

An excellent introduction and case studies can be found in Andrianakis et al. (2015), or in Vernon et al. (2010). History matching is perhaps less ambitious but correspondingly more robust than calibration methods, and a full calibration can be carried out once the history matching procedure has been completed.

McNeall et al. (2013) studied an ensemble of an ice sheet model and found that using a single type of observation for ruling out input space was not very powerful - particularly if there was not a very strong relationship between an input parameter and the simulator output. A key technique therefore is to use multiple data sets for the history matching, ruling out a candidate input space according to an empirical rule. Several rules have been used - for example using the maximum implausibility of a multiple comparison, a candidate input point point may be ruled out by a single observation. A more conservative approach is to use the second or third implausibility score, or to use a multivariate implausibility score, both introduced in Vernon et al. (2010). The aim of these scores is to ensure that an unidentified model discrepancy does not result in ruling out candidate points that are in fact perfectly good. History matching can be effective in reducing the volume of parameter space that is considered plausible to produce model runs that match the real system. For example, Williamson et al. (2015) report very large reductions (around 99

While history matching has often been used used to explore and reduce the input parameter space of expensive simulators, its use as a tool to find discrepancies, bias and inadequacies in simulators is less developed. Williamson et al. (2015) argue that what was assumed a structural bias in ocean model HadCM3 could be corrected by choosing different parameters. In a different system McNeall et al. 2016 argue that a standard set of parameters for the land surface component of the climate model FAMOUS should be retained, and that a bias seen in the simulation of the Amazon rainforest is a simulator discrepancy not a poor parameter choice.

In that case, the model simulated other forests at the standard set of parameters well, and only a tiny volume of parameter space could be found that (barely) adequately simulated all the forests. When cast as a choice between keeping the default parameters, or rejecting them and accepting the new region of parameter space, they argued that the former was more likely to produce a good model, as presumably scientific judgement and expertise informed the original choice of parameters, whereas there were a number of reasons one might reject the proposed parameter space.

1.2 Aims of the paper

A well simulated and vigorous Amazon forest at the end of the spinup phase of a simulation experiment is a prerequisite for using the model to make robust projections of future changes in the forest. The analysis of McNeall et al. 2016 (hereafter M16) identified that the land surface input spaces where FAMOUS forest fraction was consistent with observations were very different in the Amazon than they were for other forests. The area of overlap of these spaces - one that would normally be chosen in a history matching exercise - did not simulate any of the forests well, and did not contain the default parameters. M16 suggested that assuming an error in the simulation of the Amazon forest would be a parsimonious choice. Two obvious candidates for the source of the discrepancy in the Amazon were identified: (1) a lack of deep rooting in the Amazon, meaning that trees could not access water at depth as in the real Amazon and (2) a bias in the climate of the model, impacting the vigour of the trees.

This paper revisits and extends the analysis of M16 to attempt to simultaneously (1) assess the impact of a bias corrected corrected climate on the Amazon forest and (2) to identify regions of input parameter space that should be classified as plausible, given a corrected Amazon climate. To bias correct the climate we develop a new method to augment a Gaussian process emulator, with simulator outputs acting as inputs to the emulator alongside the standard input parameters. We use simulated output of forests at different geographical locations to train the emulator, describing a single relationship between the climate of the simulator, the land surface inputs and the forest fraction. In doing so, we develop a technique that might be used to bias correct existing ensembles of coupled models, allowing a more computationally efficient method for final system-tuning of models.

In section **??**, we review the literature on the possible causes of the low Amazon forest fraction in FAMOUS. In section **??**, we describe how we use the temperature and precipitation to augment the Gaussian process emulator. In section **??** we use the emulator to estimate the sensitivity of forest fraction to changes in land surface and climate parameters. In section **??** we use the augmented emulator to bias correct the climates of the forest and examine the effect of that bias correction on the input space that is deemed statistically acceptable in a history matching exercise. In section **??** we search for regions of parameter space where the bias corrected simulator might perform better than at the default parameters. In section **??** we look at regions of climate space where the default parameters would produce statistically acceptable forests. Finally, we offer some discussion of our results in section **??** and conclusions in section **??**.

2 Climate and forest fraction in FAMOUS

Previous studies have concluded that the climate state has an influence on the Amazon rainforest. Much of that work has been motivated by the apparent risk of dieback of the Amazon forest posed by a changing climate [e.g. Malhi 2008, Cox et al 2004]. We assume here that factors that might precipitate a future change in the simulated state of the Amazon rainforest might also influence the simulation of the steady-state preindustrial forest in FAMOUS. Parameter perturbations and CO2 concentrations have also been shown to influence the simulation of tropical forests in climate models, with increases in CO2 fertilisation and associated increased water use efficiency through stomatal closure offsetting the negative impacts of purely climatic changes. A metric linked to rainforest sustainability by Mahli (2009b) is Maximum Cumulative Water Deficit, which describes the most negative value of climatological water deficit measured over a year. In a similar vein Good et al. (2011, 2013) find that in Hadley Centre models, sustainable forest is linked to dry-season length, a metric which encompasses both precipitation and temperature, along with sensitivity to increasing CO2 levels. No forest is found in regions that are too warm or too dry, and there is a fairly distinct boundary between a sustainable and non-sustainable forest. Galbraith et al (2010) found that temperature, precipitation and humidity had greatly varying influences, and by different mechanisms on changes in vegetation carbon in the Amazon across a number of models, but that rising CO2 mitigated losses in biomass. Poulter et al. (2010) found that the response of the Amazon forest to climate change in the land surface model LPJml was sensitive to perturbations in parameters, but that the dynamics of a dieback in the rainforest was robust across those perturbations. In that case, the main source of uncertainty of dieback was uncertainty in climate scenario. Boulton et al. (2017) found that temperature threshold and leaf area index parameters both have an impact on the forest sustainability under projections of climate change in the Earth system version of HadCM3.

2.1 Biases in FAMOUS

M16 speculated that both local climate biases and missing or incorrect processes in the land surface model - such as missing deep rooting in the Amazon - might be the cause of the simulated low forest fraction in the Amazon region at the end of the pre-industrial period in an ensemble of the climate model FAMOUS. In this study we use the ensemble of FAMOUS previously used in M16, to attempt to find and correct the cause of persistent low forest fraction in the amazon, identified in that paper.

The Fast Met Office UK Universities Simulator, FAMOUS (Jones et al., 2005; Smith et al., 2008), is a reduced-resolution climate simulator based on the climate model HadCM3 (Gordon et al., 2000; Pope et al., 2000). The model has many features

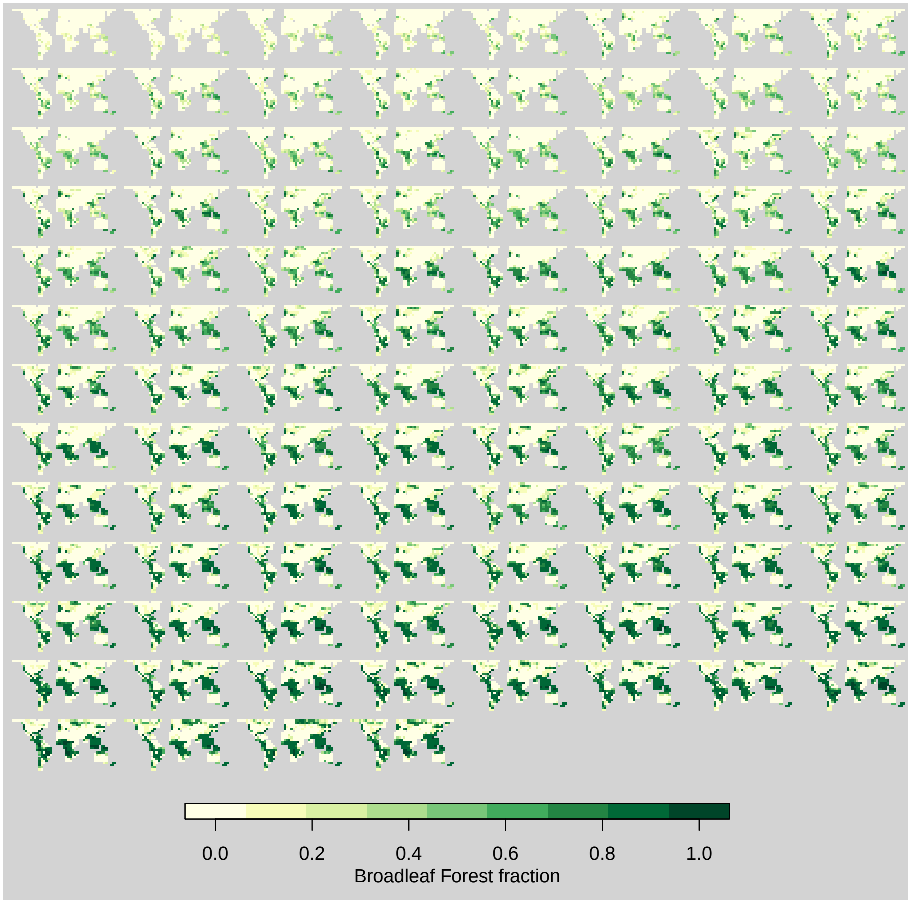

Figure 1. Broadleaf forest fraction in the FAMOUS ensemble, ranked from the smallest to largest global mean value.

of modern climate simulators, but is of sufficiently low resolution to provide fast and simple data sets with which to develop UQ methods. Full details of the ensemble can be found in M16 and Williams et al. 2013.

The ensemble of 100 members perturbed 7 land surface and vegetation inputs, which had a strong impact on vegetation cover at the end of a spinup period, with atmospheric CO_2 at preindustrial conditions (figure). The broadleaf forest fraction in individual ensemble members varies from almost non-existent to vigourous. The strong relationships between forest fraction in each forest and global values implies that perturbations in input parameters exert a larger control over all forests simultaneously, and individual forests to a smaller extent.

M16 extracted aggregated forest fraction data for the Amazon, Southeast Asian, North American and central African forests, along with the global mean. They were only able to find very few land surface parameter settings which the emulator suggested should lead to an adequate simulations of the Amazon forests and the other forests together. Further, these parameter sets were at the edges of sampled parameter space, where larger uncertainty in the emulator may have been driving the acceptance of the parameter sets.

The ensemble did however have a further perturbation - a parameter denoted “beta", which indexed into one of ten of the best-performing atmospheric parameter sets used in a previous ensemble with the same model. The beta parameter then summarised perturbations in a number of other parameters that impacted the climate of the model. Variations in the parameter did not correlate with any of the land surface parameters in the ensemble, and so was excluded from the analysis in M16.

In this study, we use the same ensemble of forest fraction data used in M16. However, we add temperature and precipitation data, present in the original ensemble but not used to build an emulator in the M16 study, to further our understanding of the causes of the low forest fraction in the Amazon region. The temperature and precipitation data summarise the effects of the parameter on the atmospheric component of the model, in a way that is directly seen by the land surface component of the model. We consider only regions dominated by tropical broadleaf forest, so as not to confound analysis by including other forests which may have a different set of responses to perturbations in parameters, rainfall and temperature.

For temperature observations we use the CRU global monthly surface temperature climatology (Jones 1999), covering the years 1960-1990 . For precipitation we use the average monthly rate of precipitation, covering the years from 1979-2001 from GPCP Version 2.2 provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at https://www.esrl.noaa.gov/psd/ (Adler et al., 2003). Vegetation fraction observations are adapted from Loveland et al. (2000), and are shown in figure 2. Although the observations all cover slightly different time periods, we expect the differences caused by harmonising the time periods to be very small compared to other uncertainties in our analysis, and to be well covered by our uncertainty estimates.

Figure 2 Observations of broadleaf forest fraction on their native grid (top), and regridded to the FAMOUS grid (bottom).

A plot of regional mean temperature and precipitation in the tropical forest regions in the FAMOUS ensemble (figure 3) indicates the form of the impact that the regional climate has on forest fraction in the climate model. Central African and Southeast Asian climates in the model simulations run in a sweep across the middle of the plot, from dry and cool to wet and warm.

It appears that a wetter climate - which would be expected to stabilize forests - broadly compensates for the forest reductions induced by a warmer climate. Within the ensemble of Central African forests for example, forest fraction increases towards the “cooler, wetter" (top left) part of the climate phase space. Beyond a certain value however, there are no simulated forests in this climatic region. It is clear from the plot that while central African and Southeast Asian forests are simulated in the large part considerably warmer than recent observations, they are also simulated considerably wetter, which might be expected to compensate forest stability. In contrast, while simulated considerably warmer, the Amazon is also slightly drier than recent observations, which might further reduce forest stability.

3 Methods

The climate model FAMOUS is computationally expensive enough that we cannot run it for a large enough number of input parameter combinations to adequately explore parameter space and find model biases. To increase computational efficiency we build a Gaussian process emulator: a statistical function that predicts the output of the model at any input, with a corresponding estimate of uncertainty. The emulator models climate model output y as a function g() of inputs x so that y = g(x). It is trained on an ensemble of model runs at set of inputs called the design matrix, denoted X, giving a sample of model output. The configuration of the design matrix is a latin hypercube (MacKay 1979), as used in e.g. Gregoire et al., 2010, Williams et al., 2013), with sample input points chosen to fill input parameter space efficiently and therefore sample relationships between input parameters effectively.

3.1 An augmented emulator

Our strategy is to augment the design matrix of input parameters X with corresponding atmospheric climate model output that might have an impact on the modelled land surface. We build an emulator that models the effects of both input parameters and climate on forest fraction. We then use the augmented emulator to bias correct each forest in turn.

We will use the emulator to describe the relationship between land surface parameters, atmospheric variables that summarise the action of hidden atmospheric parameters, and the broadleaf forest fraction. The relationships between these variables are summarised in figure 4.

We have a number of forests for each ensemble member, differing in driving influences by a different local climate. We use regional mean temperature, T, and precipitation, P, for each of the forests the Amazon, central Africa and Southeast Asia as additional inputs to augment our original design matrix of land surface parameters, X. Regional extent of each of the broadleaf forests can be found in the supplementary material.

These new inputs are outputs of the model when run at the original inputs X, and are influenced by the 10 atmospheric parameters perturbed in a previous ensemble, summarised in the “beta" parameter. We cannot control them directly and thus ensure that they lie in a latin hypercube configuration. We do however hope that they represent a wide spread of model behaviour, given wide perturbations of the input parameters.

With $n = 100$ ensemble members, we form each $n \times 1$ vector of temperature and precipitation and form an $n \times 2$ matrix of climate variables for the Amazon $C_{AZ} = [T_{AZ} P_{AZ}]$, Central Africa $C = [T_{AF} P_{AF}]$ and Southeast Asia $C_{AS} = [T_{AS} P_{AS}]$.

We use these to augment the original $n \times p$ input matrix X, creating a unique input location for each forest. We then stack these augmented input matrices together to form a single input matrix X’.

$$X' = \begin{bmatrix} X & C_{AZ} \\ X & C_{AF} \\ X & C_{AS} \end{bmatrix} \quad (1)$$

From an initial ensemble design matrix with $n = 100$ members and $p = 7$ inputs, we now have a design with $n = 300$ members and $p = 9$ inputs. Each member with a replicated set of initial input parameters (e.g members $[1, 101, 201]$), differ only in the T and P values. Figure **??** shows the composition of the resulting input matrix and output vector.

Where in M16, we built an independent emulator for each output (i.e. regional forest fraction), we now build a single emulator for all forest fractions simultaneously given input parameters, temperature and precipitation. The output vector y for the tropical forests has gone from being 3 sets of 100 values, to a single vector $[y_1, \ldots, y_n]$ of length 300.

Table 1. Mean absolute error (MAE) rounded to the first significant figure for the regular emulator, using just the seven land surface inputs, and the augmented emulator, including temperature and precipitation.

Forest	Regular emulator MAE	Augmented emulator MAE
Amazon	0.05	0.03
Southeast Asia	0.06	0.03
Central Africa	0.06	0.03
All	0.06	0.03

We model forest fraction $[y_1, \ldots, y_n]$ as a function of X' using the Gaussian process emulator of package DiceKriging (**?**) in the R statistical language and environment for statistical computing. Details of the emulator can be found in the supplementary material.

3.2 Verifying the augmented emulator

To verify that the augmented emulator adequately reproduces the simulator behaviour, we use a leave-one-out metric. For this metric, we sequentially remove one simulator run from the ensemble, train the emulator on the remaining ensemble members and predict the held-out run. We present the predicted members and the calculated uncertainty plotted against the actual ensemble values in figure **?**.

We see no reason to doubt that the emulator provides a good prediction and accurate uncertainty estimates for prediction at inputs points not yet run. We use the mean of the absolute value of the difference between the emulator prediction and corresponding held-out value to calculate the Mean Absolute Error of cross-validation prediction (MAE). Prediction error and uncertainty estimates remain approximately stationary across all tropical forests and values of forest fraction. The mean absolute error of prediction using this emulator is a little under 0.03, or 3% of the maximum possible value of the ensemble.

When compared against the regular emulator using just the land surface inputs, the augmented emulator performs well. The augmented emulator has a mean absolute error of prediction of 0.03 or 3% of the maximum possible value of the ensemble. The regular emulator built individually for each of the forests has a mean absolute error value of 0.058 - nearly double that of the augmented emulator. A breakdown of the mean absolute error of the emulator on a per-forest basis can be seen in table 1

We test the reliability of uncertainty estimates of the emulator by checking that the estimated probability distributions for held-out ensemble members match the true error distributions in the leave-one-out exercise. We create a rank histogram (see e.g. Hamill (2001) **?**) for predictions, sampling 1000 times from each Gaussian prediction distribution, and plotting the rank of the actual prediction in that distribution. The distribution of these ranks overall predictions should be uniform if the uncertainty estimates are reliable. Consistent overestimation of uncertainty will produce a peaked histogram, while systematic underestimation of uncertainty will produce a u-shaped histogram. The rank histogram produced by this set of predictions (figure 7) is close to a uniform distribution, indicating reliable predictions.

4 Analyses

4.1 Sensitivity analysis

The emulator allows us to measure the sensitivity of forest fraction to the land surface input parameters simultaneously with climate variables temperature and precipitation. We measure the one-at-a-time sensitivity to parameters and climate variables, using the emulator to predict changes in forest fraction as each variable is changed from the lowest to highest setting in turn, with all other parameters at the default settings or observed values. We present the results in figure 8. Parameters NL0 and V_CRIT_ALPHA and climate variables temperature and precipitation exert strong influences of similar magnitudes on forest fraction. Shaded regions represent the uncertainty of the sensitivity to each parameter, due to estimated emulator uncertainty of $\pm$ 2 standard deviations. This sensitivity measure does not include the extra uncertainty due to the fact that the relationships will change depending on the position of the other parameters. We do however get to see a measure of how temperature and precipitation affect the marginal response of the other parameters, as the observed climates of each forest are different. We clearly see that the response of the forest fraction to e.g. NL0 depends on climate - the forests fraction response is a noticeably different shape when varied under the mean climate of the South East Asian region.

4.2 HEADING

TEXT

4.2.1 HEADING

TEXT

5 Conclusions

TEXT

Code availability. TEXT

Data availability. TEXT

Code and data availability. TEXT

Sample availability. TEXT

Video supplement. TEXT

Appendix A

A1

Author contributions. TEXT

Competing interests. TEXT

Disclaimer. TEXT

Acknowledgements. TEXT

References

REFERENCE 1

REFERENCE 2

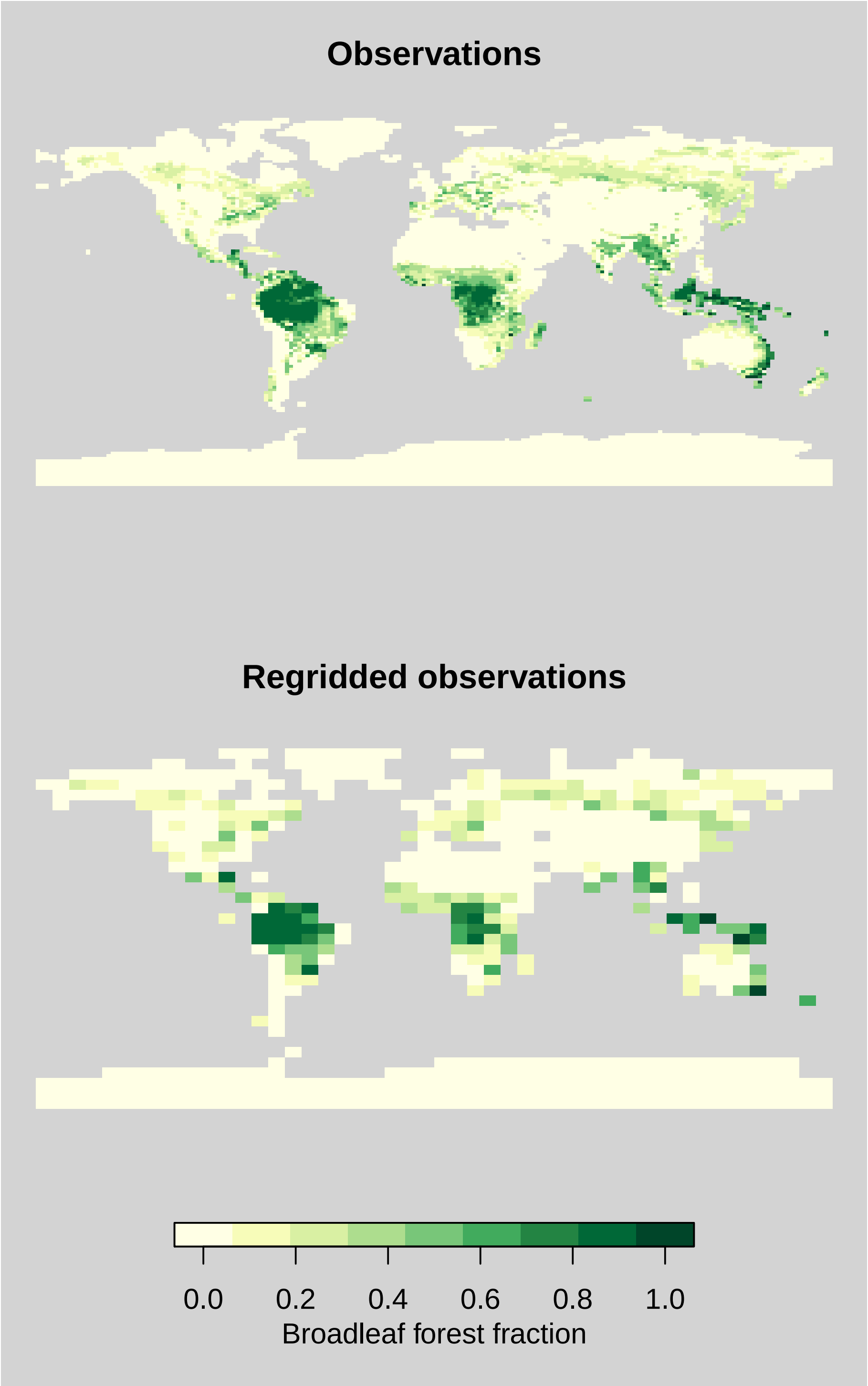

Figure 2. Observations of broadleaf forest fraction on their native grid (top), and regridded to the FAMOUS grid (bottom)

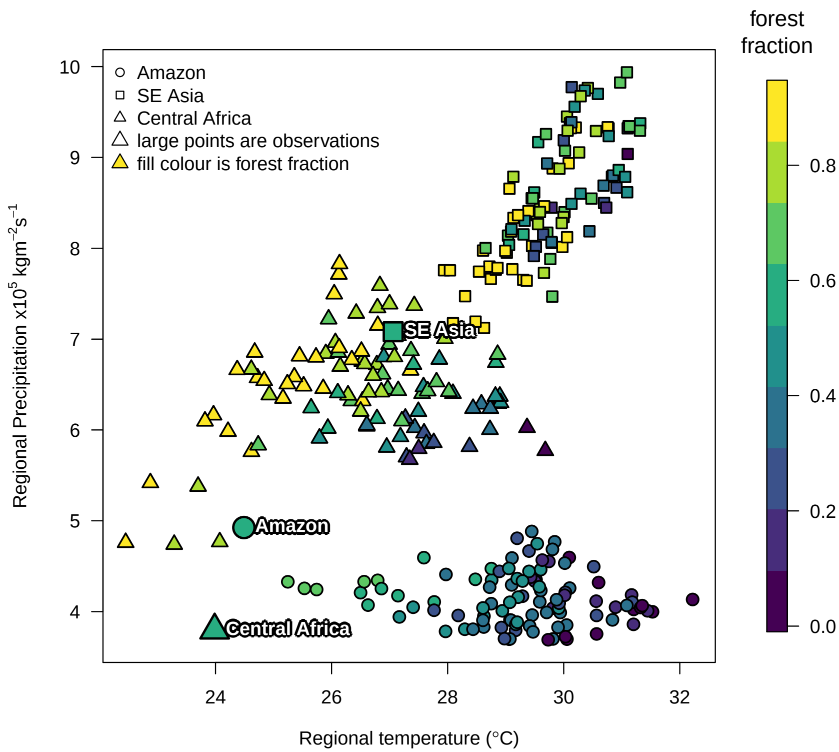

Figure 3. Smaller symbols represent broadleaf forest fraction in the FAMOUS ensemble against regional mean temperature and precipitation. Ensemble member forest fraction in the Amazon is represented by the colour of the circles, Central Africa by triangles and SE Asia by squares. Larger symbols represent observed climate and forest fraction.

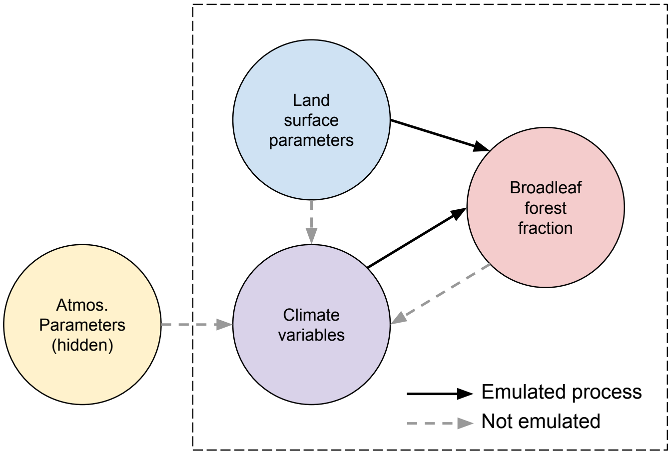

Figure 4. A graph showing the assumed relationship between input parameters, climate variables and forest fraction. An arrow indicates influence in the direction of the arrow. Processes that are directly emulated are shown with a solid arrow, while the processes shown by a dotted arrow are not directly emulated.

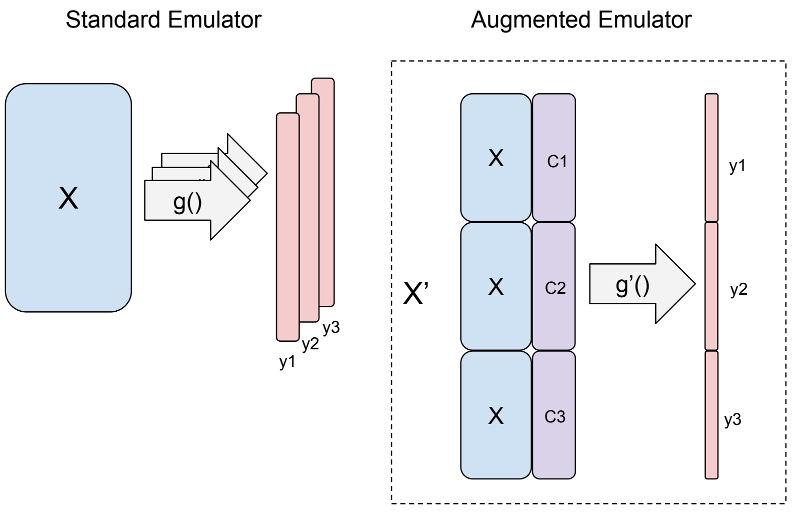

Figure 5. In a standard emulator setup (left), training data consists of an input matrix X and corresponding simulator output y. A new emulator $g_1, \ldots, ()g_n()$ is trained for each output $y_1, \ldots, y_n$ of interest. In the augmented emulator, output from the simulator $C_1, \ldots, C_3$ augments the design matrix, with the initial inputs X repeated.

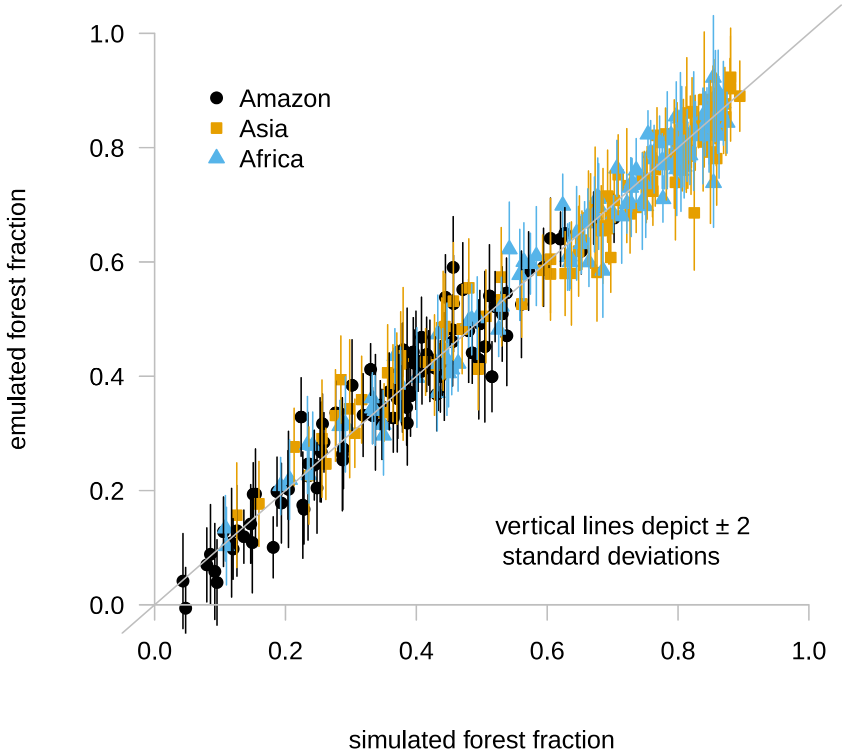

Figure 6. Leave-one-out cross validation plot, with the true value of the simulator output on the x-axis, and predicted output on the y-axis. Vertical lines indicate ± 2 standard deviations.

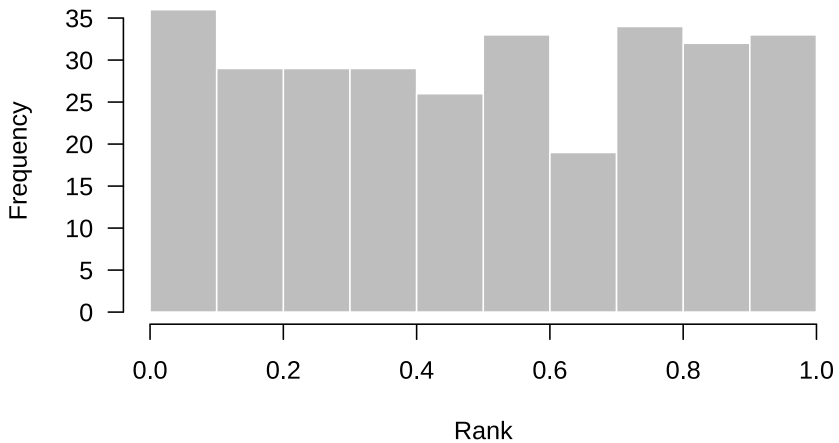

Figure 7. Rank histogram of leave-one-out predictions. For each prediction of a held-out ensemble member, we sample 1000 points from the Gaussian prediction distribution, and then record where the true held-out ensemble member ranks in that distribution. We plot a histogram of the ranks for all 300 ensemble members. A uniform distribution of ranks indicates that uncertainty estimates of the emulator are well calibrated.

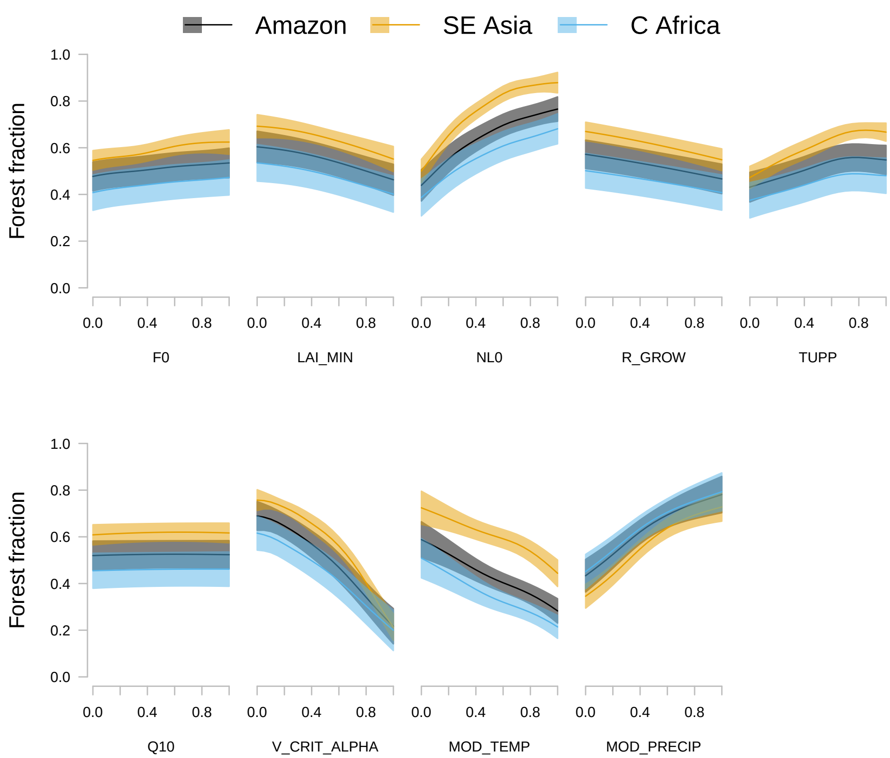

Figure 8. One-at-a-time sensitivity of forest fraction variation of each parameter and climate variable in turn across the entire ensemble range. All other parameters or variables are held at their default values while each parameter is varied. Solid lines represent the emulator mean and shaded areas represent $\pm$ 2 standard deviations of emulator uncertainty.