TEXT

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

[1]Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB UK
[2]NIWA, New Zealand
[3]University of Exeter, North Park Road, Exeter EX4 4QE UK

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

Abstract. 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 computationally cheap statistical proxy (emulator) of the climate model, we compare an ensemble of simulations of forest fraction with observations, and rule out parameter settings where the forests are poorly simulated.

Regions of parameter space where FAMOUS best simulates the Amazon forest fraction are incompatible with the regions where FAMOUS best simulates other forests. Previous studies using climate models have used similar methods to find previously untried candidate input parameter sets that remove what was assumed an underlying structural error. We offer a counter example, arguing that we have found a true structural discrepancy. This has implications for the calibration of FAMOUS: using observations of different forest regions to calibrate the model leads to very different conclusions about the best values, the corresponding uncertainty of input parameters, and potentially, predictions of future forest cover. Dealing with this structural discrepancy is vital when choosing a set of "best" parameters for the land surface - failure to do so could lead to poor parameter selection.

We characterise the structural model discrepancy, and explore the consequences of ignoring it in a history matching exercise. We perform a sensitivity analysis to find the parameters most responsible for simulator error and therefore most promising for tuning. 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 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.

Finally, we perform a history matching exercise using credible estimates for simulator discrepancy and observational uncertainty terms. We find that we are unable to constrain the parameters individually, but that 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 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 value of the any parameter coefficient that should be used to best represent the system being simulated. 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 can be represented by using a range of values for each of the coefficients in an ensemble of simulator runs.

Choosing appropriate values of these coefficients is a major research effort that encompasses domain specific, statistical and computational literature. The coefficients are tuneable by comparison of the behaviour of the simulator with observations of the real system, although there may also be direct measurements of the value of the coefficient or other 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 formal probabilistic setting, then it may also provide probability distributions for the parameters of the model, and is known as *calibration*. The process of choosing a single best parameter set is often called *tuning*. *History matching* provides a formal way of ruling out parameter settings that are inconsistent with observed data.

The motivation for calibration of a simulator is twofold. First, a simulator which matches the underlying dynamics of a system well will produce more accurate predictions. Second, a more tightly constrained parameter set will provide a narrower range of uncertainty in future predictions.

1.1 Calibration of Land surface components

Parametric uncertainty in the land surface and carbon cycle component of models is expected to represent a large fraction of current uncertainty in future climate projections ((Booth et al., 2012), (Booth et al., 2013), (Huntingford et al., 2009)). These components have been introduced into climate models 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 reducing future land carbon cycle uncertainty by identifying which parts of possible model parameter space are inconsistent with observed properties of the real climate system.

There is also a long history of statistical and data assimilation approaches used to constrain process model parameters. In the land surface model context these extend back to (Sellers et al., 1996). Recent examples are community efforts to develop a systematic set of observations to benchmark land surface processes against metrics of real world processes, for example the International Land Model Benchmarking Project (Luo et al., 2012), and PALS (Abramowitz, 2012). Such benchmarks involve an extensive set of metrics, covering a broad cross-section of model processes. These benchmarks enable an assessment of overall model skill and highlight particular areas where the model falls short. They provide a useful framework to assess improvements in model skill that arise from continual model development as well as prioritising resources towards model processes that are less well simulated. The large number of observed metrics for diverse aspects of the model processes also help avoid model parameters being tuned to address a particular process, to the detriment of wider model performance. One of the limitations of the benchmarking approach is that there is only limited current understanding of what information a given observed metric implies about the model formulation or parameters, or what this might imply about future projected changes.

1.2 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 be known as model (or simulator) bias, model error, or structural error. A useful definition from (Williamson et al., 2014) is that *A climate model bias [simulator discrepancy] represents a structural error if that bias cannot be removed by changing the parameters without introducing more serious biases to the model*. One of the main aims of the model development process is to efficiently identify important simulator discrepancies and correct them, or allow them to be taken into account in analyses; for example, during prediction using the simulator.

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, has a predictable effect on simulator output. Alternatively, the discrepancy might be due to some missing and unknown process in the model. This sort of discrepancy might appear as a bias, and only become 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.

(Kennedy and O'Hagan, 2001) introduced a Bayesian framework to 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 discrepancy into account in calibration can lead to overconfident and inaccurate estimates of the parameters, and consequently the predictions of the model (e.g. (Brynjarsdóttir and O'Hagan, 2014), (Higdon et al., 2008)). Further, even inadequate (as opposed to outright wrong) specification of a simulator discrepancy can lead to overconfidence and bias in parameters and predictions.

1.3 Paper aims and outline

Our aim is to identify parameter sets for 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 discrepancy, before performing a second history match with an appropriate discrepancy function.

In section **??** we briefly describe the ensemble of a climate simulator, an emulator and the history matching technique that we use to explore simulator discrepancy. We perform an initial history matching exercise in section **??**. We use the emulator to quantify the relationships between the simulated forest fraction and a set of model input parameters in a sensitivity analysis in section **??**. Next, we measure the performance of the model ensemble in simulating forest fraction in section **??**. We see how much input space would be ruled out as implausible in various scenarios of data combination and uncertainty budget in **??** and we learn what each individual observation tells us about input space in section **??**. In section **??**, we use the emulator and an implausibility measure to find the "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 in section **??**. In section **??**, we discuss the consequences of our findings for models of the Amazon rainforest. We offer conclusions in section **??**.

2 Data and Methods

2.1 The FAMOUS climate model

We use a pre-existing ensemble of the climate model FAMOUS throughout this study. The Fast Met Office UK Universities Simulator (FAMOUS, (Jones et al., 2005), (Smith et al., 2008)) is a reduced resolution climate simulator, based on, and tuned to replicate, the climate model HadCM3 ((Gordon et al., 2000); (Pope et al., 2000)). Computational efficiency is gained primarily through reduced resolution. Atmospheric grid boxes are 4 times the size of HadCM3, and ocean 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 (Cox, 2001). FAMOUS runs approximately 10 times faster than HadCM3, making it ideal for running large ensembles, or long integrations, with modest supercomputing facilities.

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

The explicit inclusion of vegetation in FAMOUS is documented in (Williams et al., 2013), which introduces surface tiling in the newer MOSES2 scheme. Five different vegetation types are simulated: broadleaf and needleleaf trees, C3 and C4 grasses, and shrubs. Each of these has a fractional coverage in each gridbox. Several surface types represent the absence of vegetation: bare soil, land ice, urbanised land use and inland water. (Williams et al., 2013) describe the optimisation of carbon cycle parameters in both the terrestrial and ocean domains, validated against available observations and reanalysis products. After this optimisation procedure, climatologies are presented using both fixed and dynamic vegetation, where the surface types the ensemble, with respect to the corresponding observations.

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 regimes relevant for the Amazon rainforest. (Smith et al., 2008) noted: "the Amazon region is not wet enough for a fully humid region to exist". These regimes are formulated using the Koppen-Geiger climate classification scheme, which is based on temporal and magnitude distributions of precipitation and temperature (Gnanadesikan and Stouffer, 2006).

2.3 The ensemble

We use an ensemble of 100 simulations of FAMOUS detailed in (Williams et al., 2013), and build upon the results of that study. The ensemble was run in order to test the utility of including the carbon cycle in enhancing the FAMOUS model. The ensemble design perturbs 7 vegetation and land surface control parameters (see table **??**) in a latin hypercube configuration (McKay et al., 1979). This kind of design efficiently spans parameter space, and has been shown to be better than others for constructing surface response type statistical models known as emulators (Urban and Fricker, 2010).

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

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

The ensemble simulates the 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 CO2. The vegetation dynamics component of the simulator, TRIFFID, is run in "equilibrium" mode to increase computational efficiency. Here, the vegetation dynamics module is run for the equivalent of 10,000 years for each decade of climate. 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 adapted from (Loveland et al., 2000), and are regionally aggregated versions of the data used in the previous study by (Williams et al., 2013). We use broadleaf only for the tropical forest, and a mixture of broadleaf and needleleaf for the North American forest. A summary of the forest fraction data in the ensemble can be found in figures 1 and **??**. The former shows the spatial distribution of forest fraction in FAMOUS, showing maps of both the mean and standard distribution across the ensemble of 100 members. The parameter ranges are not explicitly chosen to 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 Africa, South East Asia, North America and Global regions (see supplementary information for region details).

South East Asia 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 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 VCRITALPHA (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, is not fast enough to evaluate at every viable candidate point within input space, termed $\mathcal{X}$. We therefore use a computationally cheap statistical proxy to the simulator, called an emulator. The emulator provides a prediction of simulator output at any required untested input, many 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 extra uncertainty term to account for the fact that the emulator is not a perfect prediction of the simulator output. We use a gaussian process emulator that assume zero uncertainty at points where the model has already been evaluated, 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 of the simulator FAMOUS y as a deterministic function of a vector of input parameters x, such that $y = g(x)$. The emulator is a nonlinear regression model conditioned on a sample $Y = g(X)$, and provides a prediction of simulator output, $\hat{y}$ at untested input sets $x \in \mathcal{X}$ such that

$$\hat{y} = \eta(X) + e \tag{1}$$

where η represents the emulator, as a substitute for $g()$, and e represents the uncertainty necessary because the emulator is not a perfect prediction. We build a number of emulators of the ensemble, the details for each depending on the application. All use the DiceKriging package (Roustant et al., 2012), in the statistical programming environment (R Core Team, 2016).

DiceKriging allows the user flexibility in specifying the emulator, and then estimates hyperparameters using the training data. We use "Universal Kriging" and specify a linear prior for each emulator. This means that the emulator starts with a linear model, and then models deviations from this as a Gaussian process. 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 material).

2.6 History matching

We aim to repeat the achievement of (Williamson et al., 2014), to 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 finding a set of input parameters $x*$ where the output of the model is deemed tolerably close to the observations, given uncertainty in the observations and known deficiencies of the model. Constraining parameters in this way should help 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 "Not Ruled Out Yet" (NROY). As such, we regard them as conditionally accepted, contingent on new observations or information. History matching was developed by (Craig et al., 1997), and has been used extensively in hydrocarbon extraction sciences, and astronomy (e.g. (Vernon et al., 2010)). It has been used to confront climate simulators with observations, for example by (Lee et al., 2016), (Williamson et al., 2013), (Ritz et al., 2015),]. The potential of an observational dataset to constrain NROY input space depends on the strength of the relationship between the inputs and simulator outputs, the accuracy of the model, the uncertainty of the observations, and the tolerance of the modeller towards a mismatch between the model and the observations (McNeall et al., 2013).

Observations of the system are denoted z, and we assume that they are made with uncorrelated and independent errors η such that $z = y + \eta$. Assuming that the simulator contains a systematic structural discrepancy δ, then the observations can be related to the input parameters

$$z = g(x*) + \delta + \eta \tag{2}$$

If the simulator were fast enough to evaluate at 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, is not fast enough for this. 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 proxy for the model output, replacing $g(x)$ with $\eta(x)$ in equation 2, and including a term for emulator uncertainty in the history matching calculations.

Each point in input space is assigned an Implausibility I, according to equation **??**. 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 uncertainty about the observation , or uncertainty about the model discrepancy , or the emulated output at that input .

$$I^2 = |z - E[g(x)]|^2 / var(g(x)) + var(\delta) + var(\eta) \tag{3}$$

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 within 3 standard deviations of the mean (Pukelsheim, 1994).

Any input parameter set that has an implausibility score below the threshold is designated ‘Not Ruled Out Yet’ (NROY), and is retained for further analysis. It should be noted that this does not imply that the input setting is *good* merely that the evidence from observations is not sufficient to rule it out as implausible: this may change as more observations, or more simulator runs become available.

(Craig et al., 1997) (Booth et al., 2012) (Booth et al., 2013) (Huntingford et al., 2009) (Sellers et al., 1996) (Abramowitz, 2012) (Luo et al., 2012) (Williamson et al., 2014) (Kennedy and O’Hagan, 2001) (Brynjarsdóttir and O’Hagan, 2014) (Higdon et al., 2008) (Jones et al., 2005) (Smith et al., 2008) (Gordon et al., 2000) (Pope et al., 2000) (Cox, 2001) (Smith, 2012) (Williams et al., 2013) (Williams et al., 2014) (Gnanadesikan and Stouffer, 2006) (McKay et al., 1979) (Urban and Fricker, 2010) (Gregoire et al., 2010) (Loveland et al., 2000) (Roustant et al., 2012) (R Core Team, 2016) (Vernon et al., 2010) (Lee et al., 2016) (Williamson et al., 2013) (Ritz et al., 2015) (McNeall et al., 2013) (Pukelsheim, 1994) (Carslaw et al., 2013) (Saltelli et al., 1999) (Pujol et al., 2015) (Cox et al., 2004) (Good et al., 2008) (Joetzjer et al., 2013) (Staver et al., 2011) (Malhi et al., 2009) (Yin et al., 2012)

3 HEADING

TEXT

3.1 HEADING

TEXT

3.1.1 HEADING

TEXT

4 Conclusions

TEXT

Appendix A

A1

Author contributions. TEXT

Acknowledgements. TEXT

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.

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.

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.

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.

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.

Ritz, C., Edwards, T. L., Durand, G., Payne, A. J., Peyaud, V., and Hindmarsh, R. C.: Potential sea-level rise from Antarctic ice-sheet instability constrained by observations, Nature, 528, 115–118, 2015.

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.

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.

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.

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.

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.

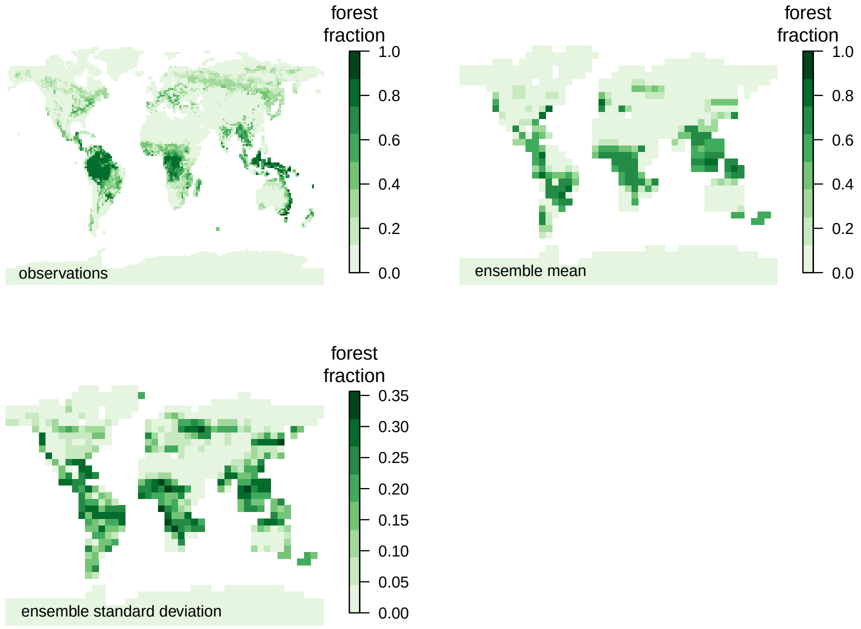

Figure 1. TEXT

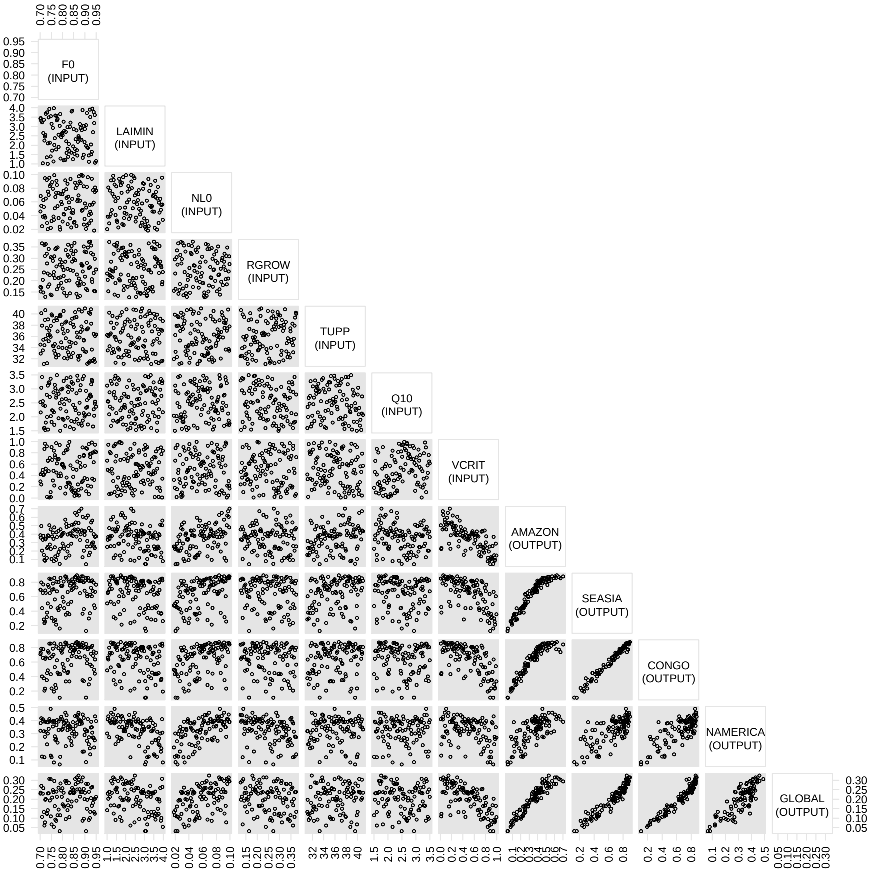

Figure 2. TEXT

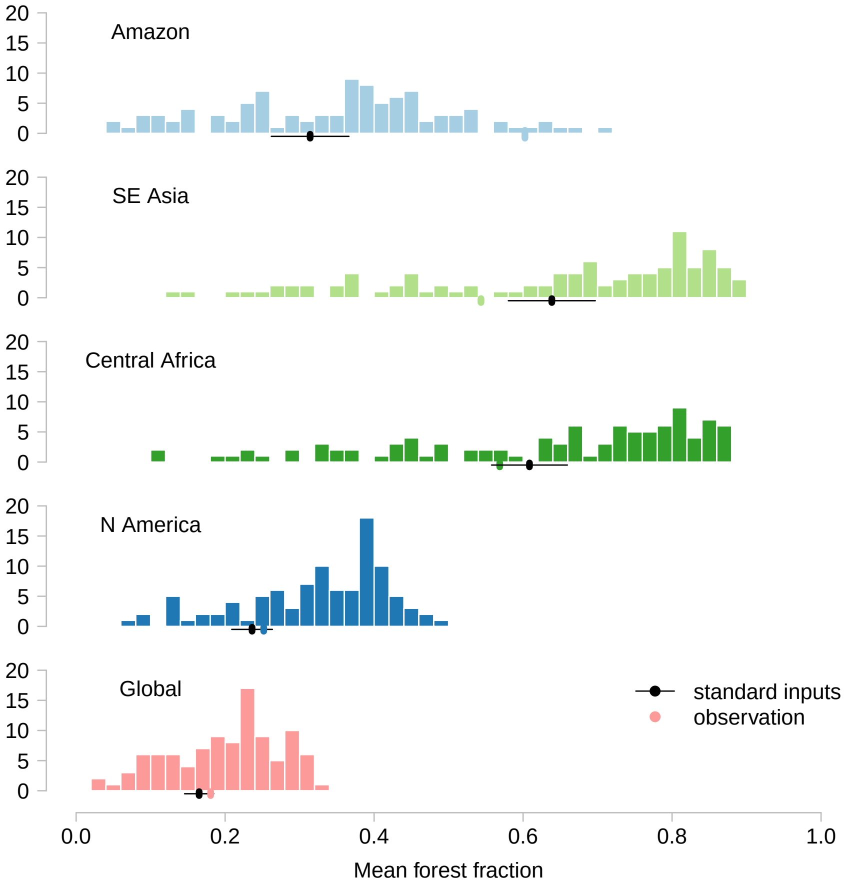

Figure 3. TEXT

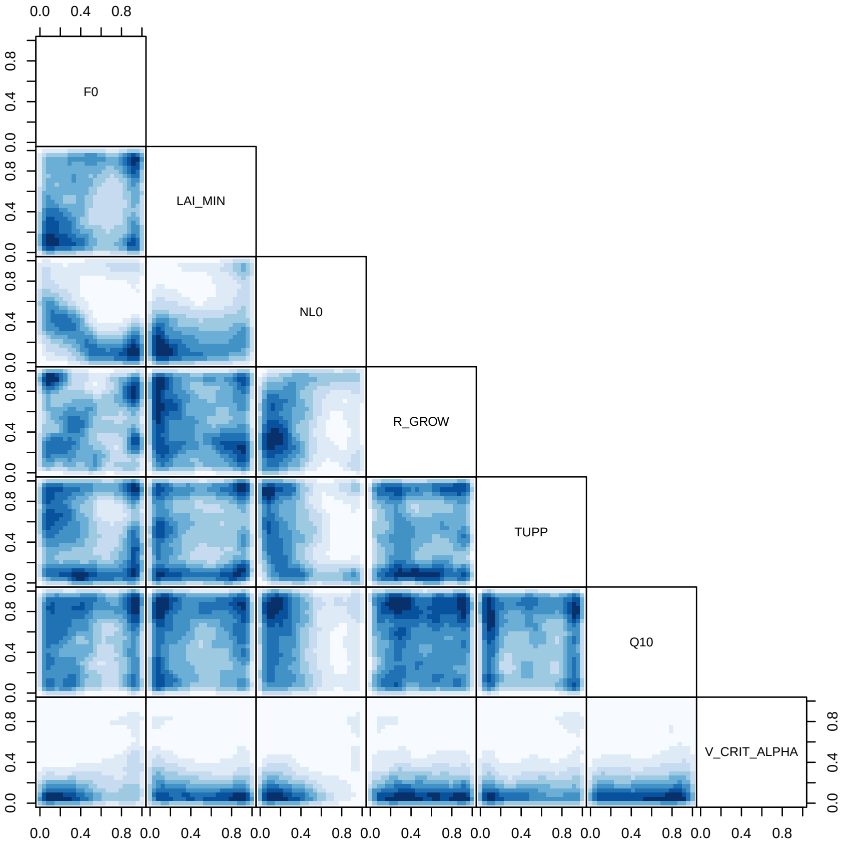

Figure 4. TEXT

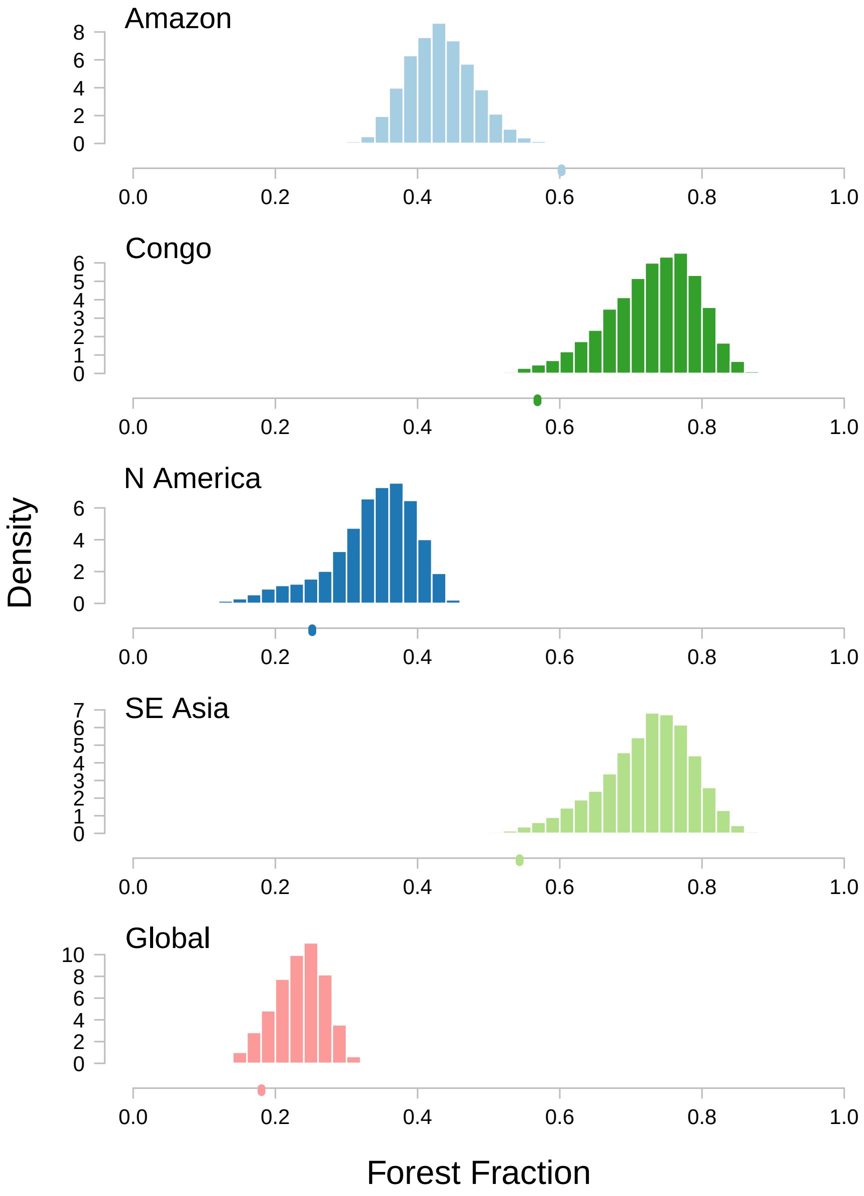

Figure 5. TEXT

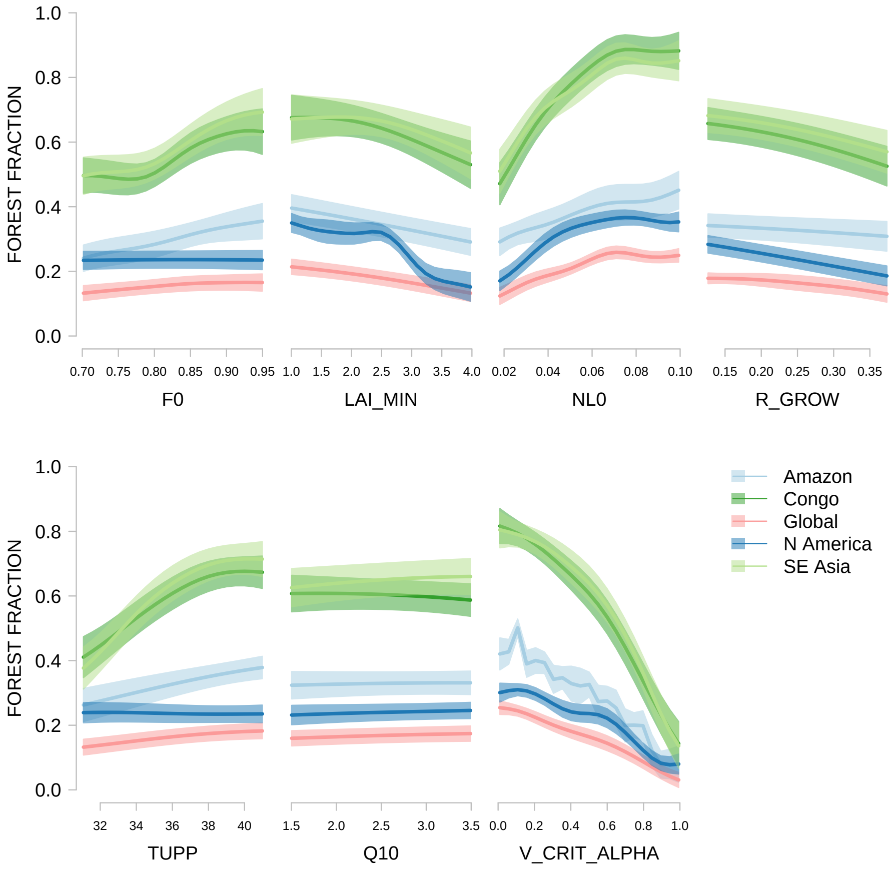

Figure 6. TEXT

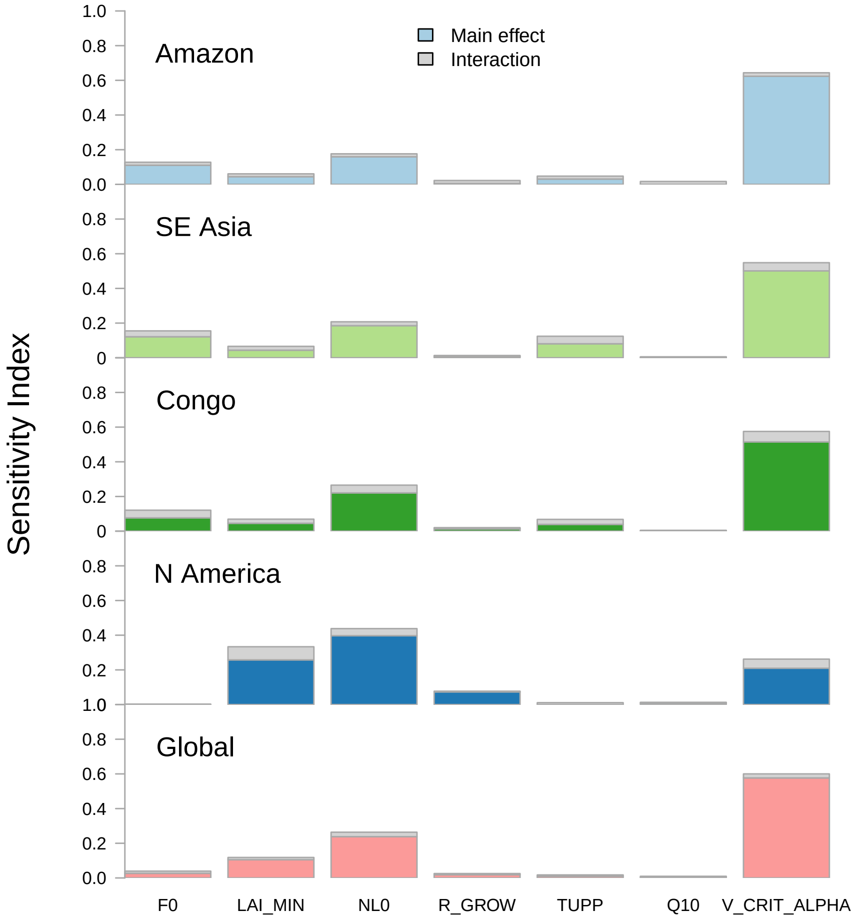

Figure 7. TEXT

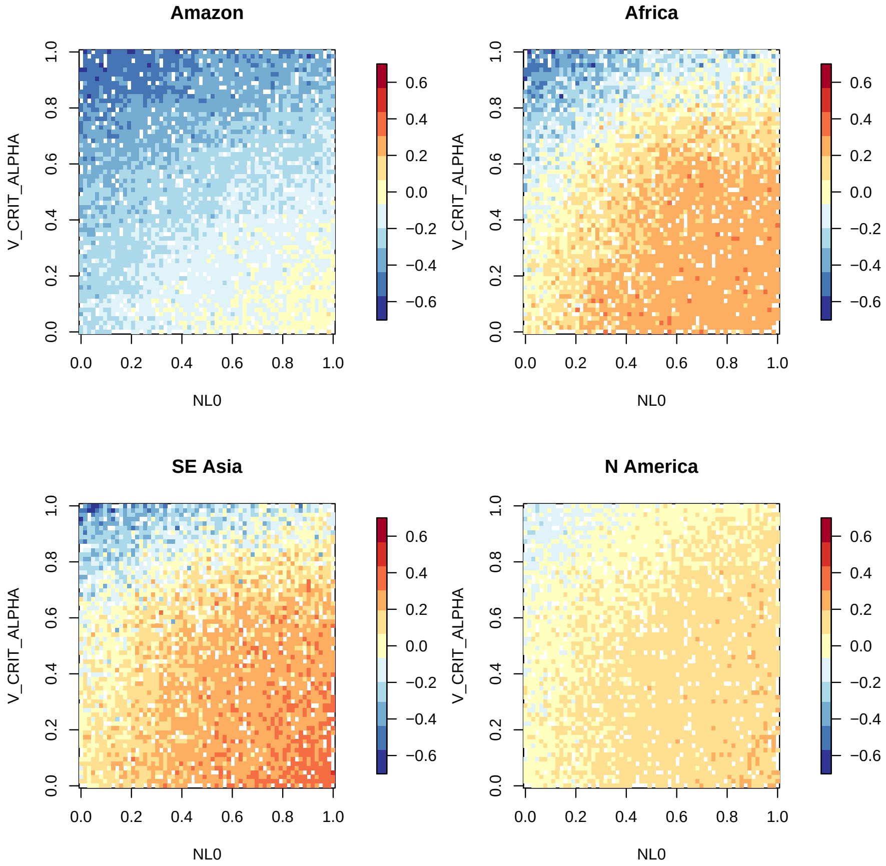

Figure 8. TEXT

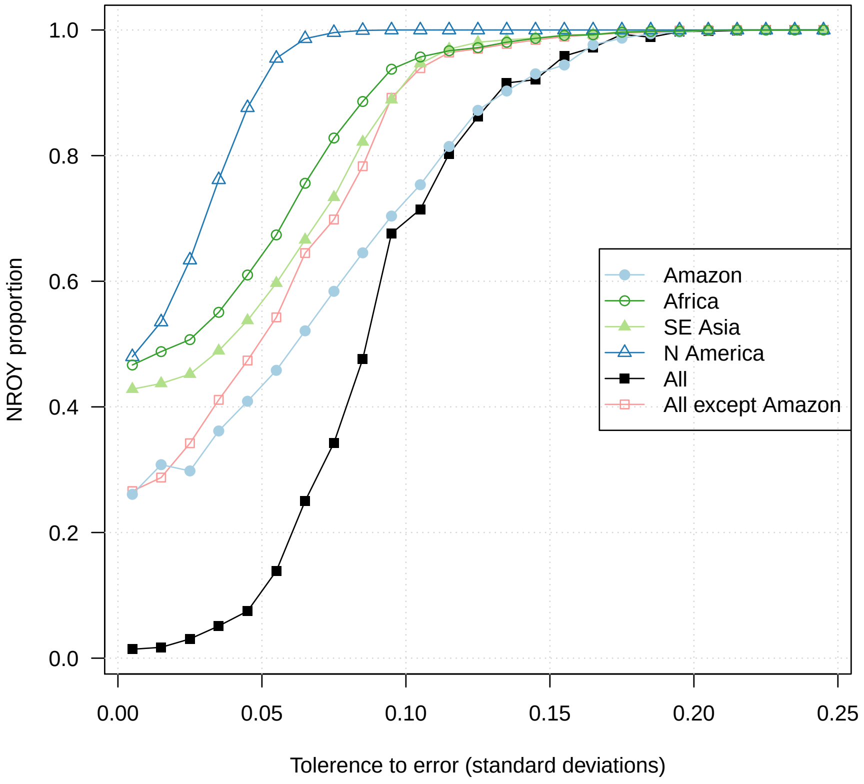

Figure 9. TEXT

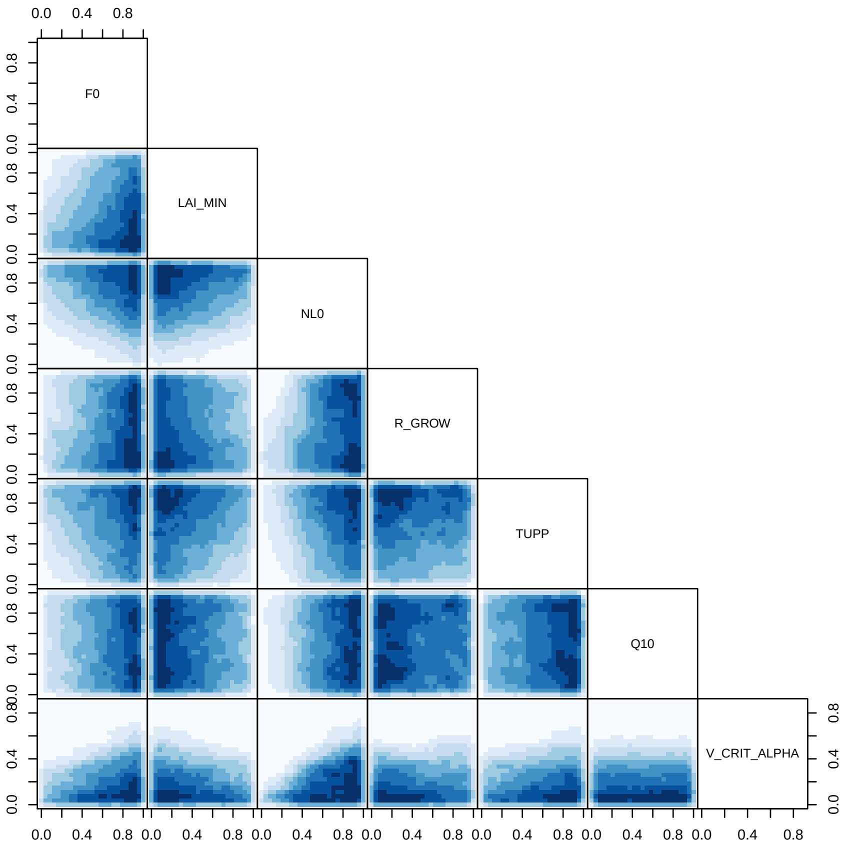

Figure 10. TEXT

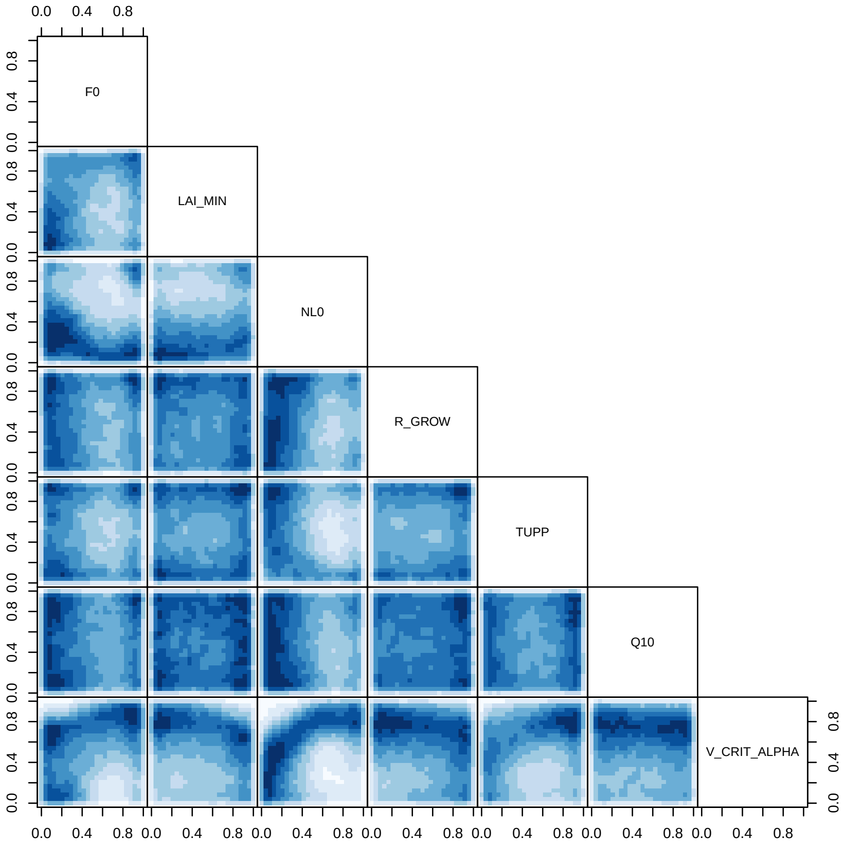

Figure 11. TEXT

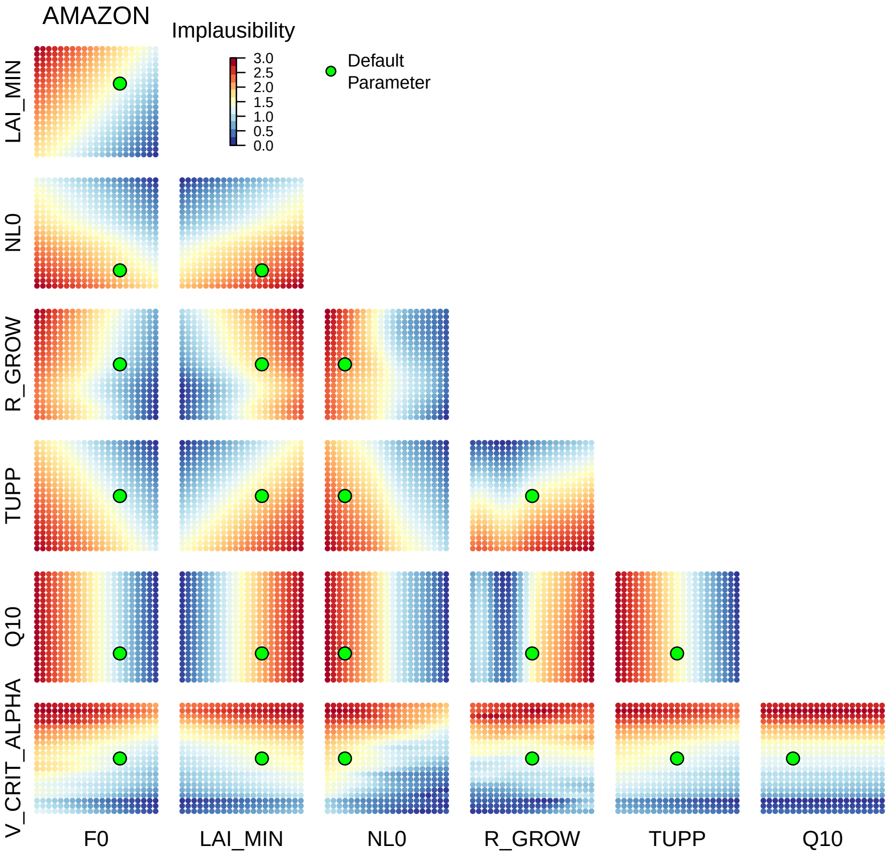

Figure 12. TEXT

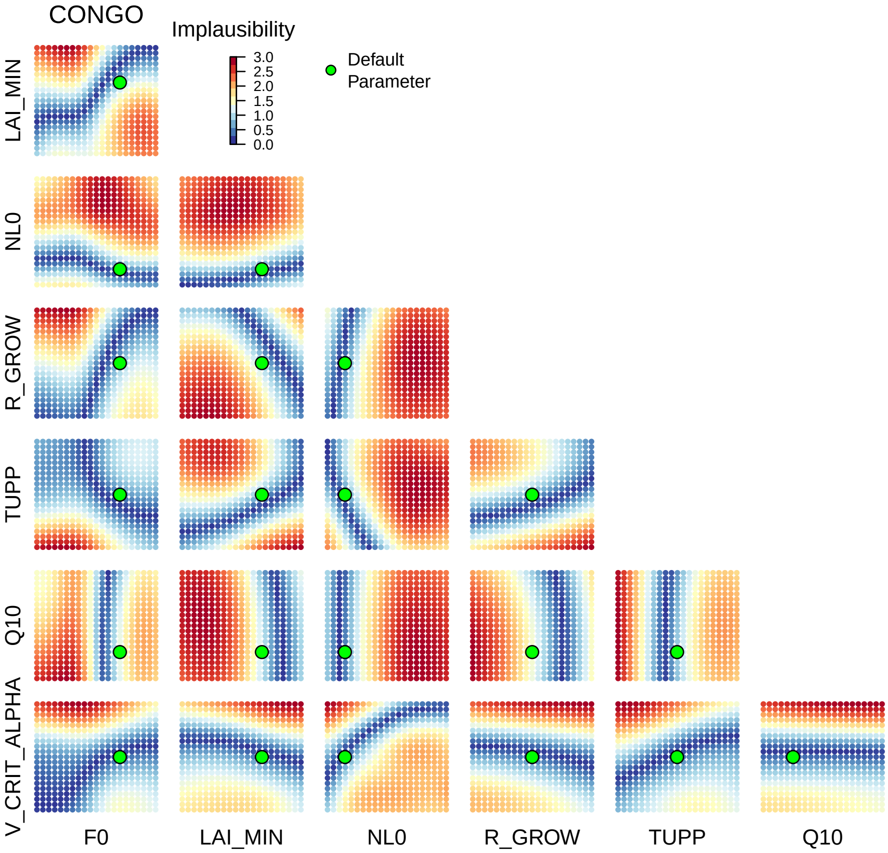

Figure 13. TEXT

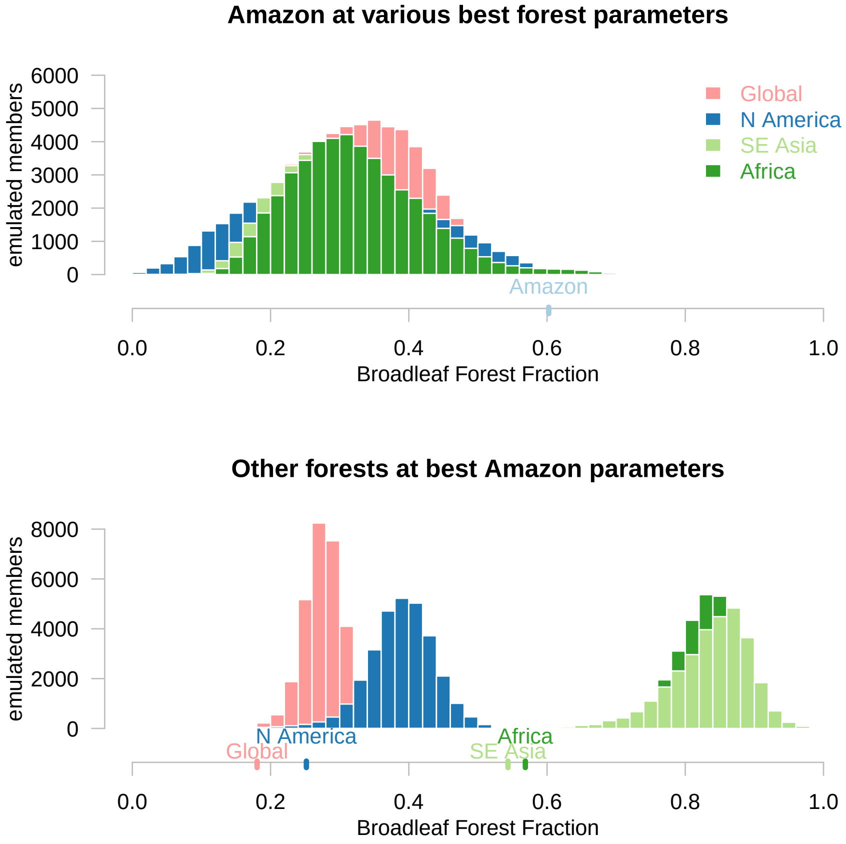

Figure 14. TEXT

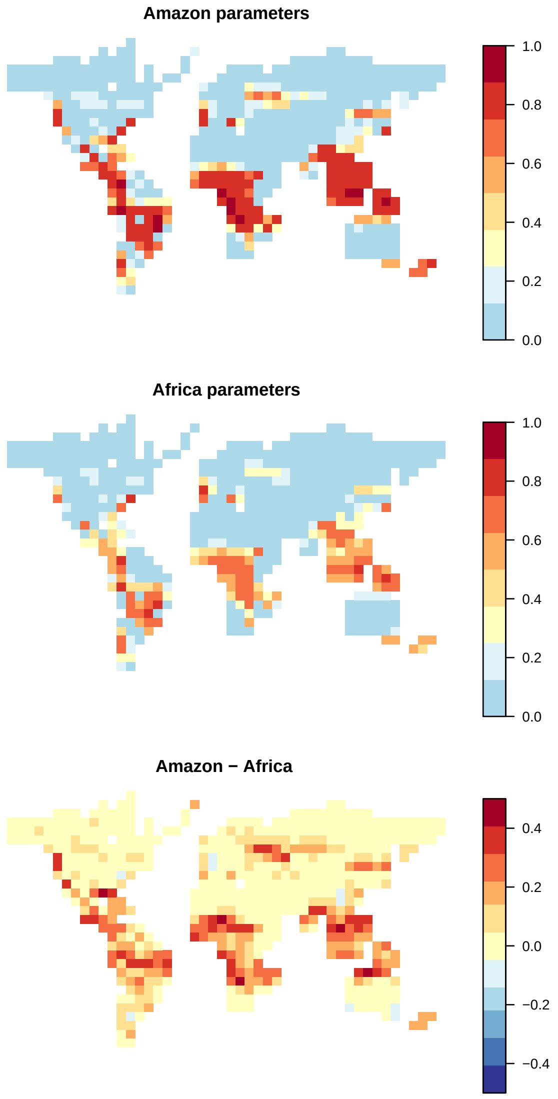

Figure 15. TEXT

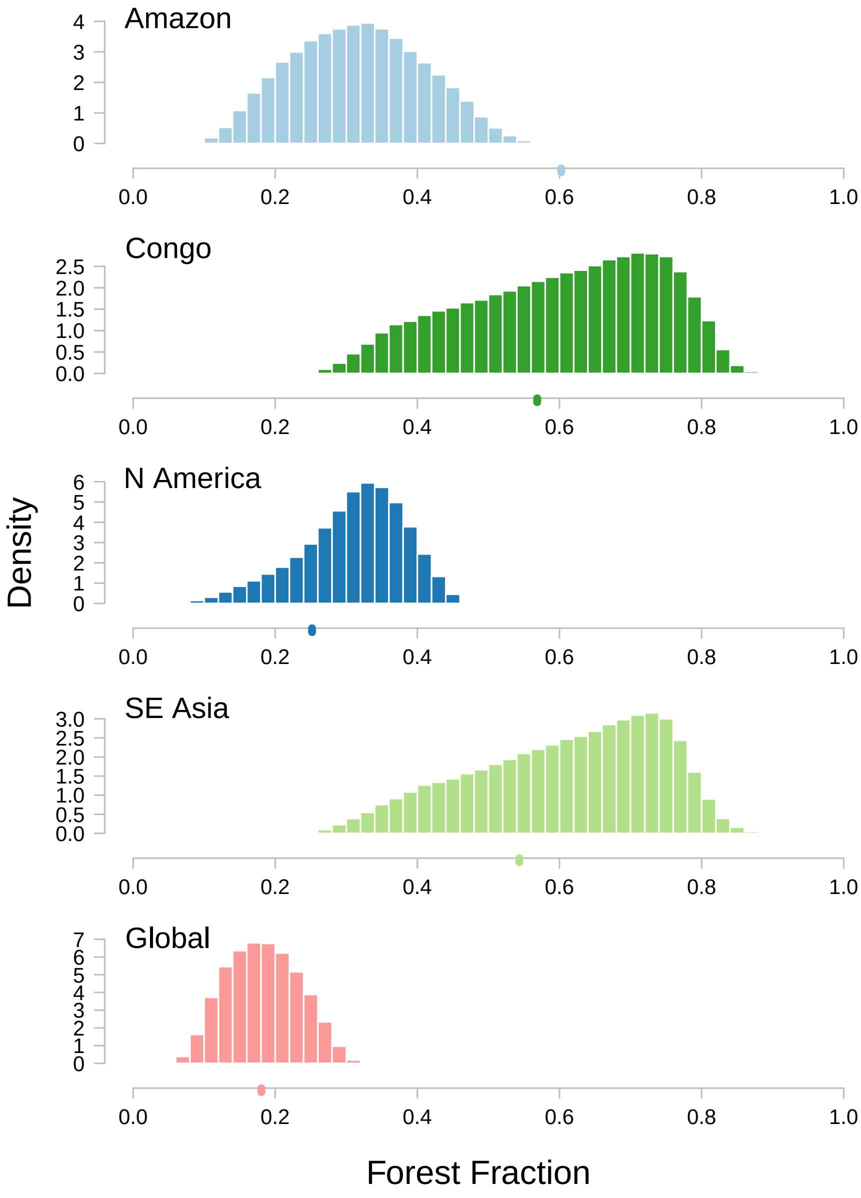

Figure 16. TEXT

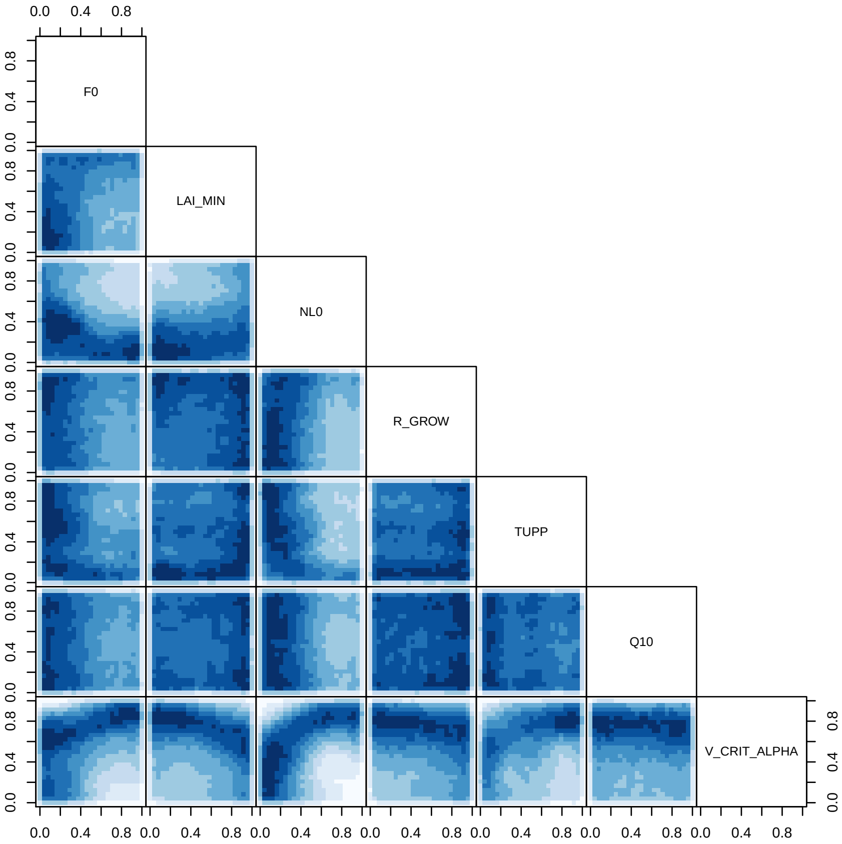

Figure 17. TEXT

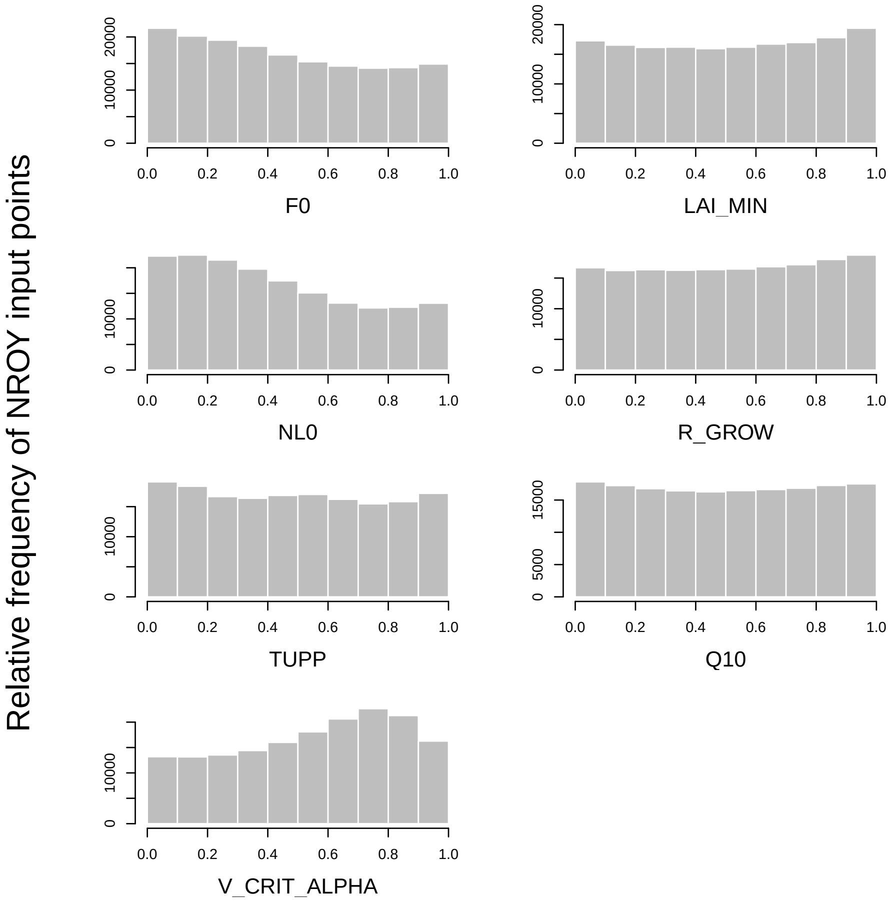

Figure 18. TEXT