Emulation of high-resolution land surface models with sparse Gaussian processes with application to JULES

Evan Baker[1], Anna Harper[2], Daniel Williamson[2], and Peter Challenor[2]

[1]Living Systems Institute, University of Exeter, Exeter EX4 4QD, UK
[2]Department of Mathematical Sciences, University of Exeter, Exeter EX4 4QF, UK

Correspondence: Evan Baker (e.baker@exeter.ac.uk)

Abstract.

Land surface models are typically integrated into global climate projections, but as their spatial resolution increases the prospect of using them to aid in local policy decisions becomes more appealing. If these complex models are to be used to make local decisions, then a full quantification of uncertainty is necessary, but the computational cost of running just one

5   simulation at high resolution can hinder proper analysis.

Statistical emulation is an increasingly common technique for developing fast approximate models in a way that maintains accuracy but also provides comprehensive uncertainty bounds for the approximation. In this work, we develop a statistical emulation framework for land surface models which acknowledges the forcing data fed into the model, providing predictions at a high resolution. We use The Joint UK Land Environment Simulator (JULES) as a case study for this strategy, and perform

10   initial sensitivity analysis and parameter tuning to showcase its capabilities. JULES is perhaps one of the most complex land surface models, and so our success here suggests incredible gains can be made for all types of land surface model.

1 Introduction

Land surface models (LSMs) represent the terrestrial biosphere within weather and climate models, focusing on hydrometeo-rology and biogeophysical coupling with the atmosphere. The latter includes nutrient flows between vegetation and soils, and

15   the turbulent exchange of CO2, heat, moisture, and momentum between the land surface and the atmosphere. The Joint UK Land Environment Simulator (JULES, Cox et al. 1998, Best et al. (2011), Clark et al. (2011)) is an example of an LSM; used for a variety of applications and temporal/spatial scales as part of the UK Met Office's Unified Modelling system. LSMs can be used to further scientific understanding of land surface processes and to inform policy decisions. For both applications, increased confidence in simulated results and knowledge of model uncertainty is needed, which typically involves running the

20   model many times with varied forcings and parameters (Booth et al., 2012; Murphy et al., 2004). The computational cost of running these models limits the number of runs that can be obtained, constraining the resulting analysis.

An important factor in computational cost versus practical relevance is the resolution at which the model can be run. Whilst there is a general, and not always justifiable, push towards higher resolution across climate modelling, when using LSMs to support landscape decisions (whether these be local/national government policy decisions, or landowner investment decisions),

25   ensuring that the model is able to inform at decision relevant resolutions is critical. For example, if considering different policies to incentivise farmers to alter land use (by giving over land to tree planting for example), LSM simulations run to help understand the efficacy and risks of different policies would need to be at a high enough resolution to capture areas the size of individual farms (say at 1km). For example, in Ritchie et al. (2019), at 1.5km x 1.5km resolution for Great Britain (77980 grid cells), JULES took 30 hours to simulate 11 years (with 20 years of spin up simulation to allow any input parameters to

30   influence the present day land surface, which took 1 additional hour) on 144 cores of the UK NERC/Met Office MONSooN supercomputer. Quantifying uncertainty for even a single policy option, let alone a diverse array of policies, would not be feasible using high resolution JULES directly.

Statistical surrogate models, also known as emulators, have been developed to combat this issue (Sacks et al., 1989; Kennedy and O'Hagan, 2001). An emulator is a statistical model that, once built, facilitates fast predictions of the output of a computer

35   model, with uncertainty in the prediction, and without any further simulation. The resulting statistical model provides a powerful tool for exploring, understanding, and improving the process-based model from which it is built. Fast predictions of land surface models could enable better decision making, improve scientific understanding, and enable the effective linking of multiple models, all while quantifying the various uncertainties involved.

Emulators have been widely used by the climate community. They have been used to study the Met Office's coupled models

40   (Williamson et al., 2013), including developing the UK Climate Projections in 2009 and 2018 (Sexton et al., 2012, 2019). Hemmings et al. (2015) build mechanistic emulators for specific locations for an ocean biogeochemical model; Petropoulos et al. (2014) conduct a comprehensive sensitivity analysis using an emulator for a land surface model; and Williamson et al. (2017) calibrates an ocean model using emulators. Such efforts typically only emulate a few key locations or summary statistics of interest, for example; McNeall et al. (2020) emulate (and calibrate) JULES considering only three averaged locations. These

45   types of analysis are useful for understanding the sensitivity of the model output to its different parameters, and for constraining parameter space, yet cannot be used as surrogates to the full model to support decision making.

Emulating the spatio-temporal output in order to use the emulator as a surrogate to the model, has also received a lot of attention. Lu and Ricciuto (2019) attempt to emulate an LSM at a higher resolution, where they reduce the dimensionality of the output via singular value decomposition. This is a well-known strategy (Higdon et al., 2008), but lowering the dimension in this

50   way can lead to a loss of information and interpretability, with documented negative effects (Salter et al., 2019). Additionally, Lu and Ricciuto (2019) use a neural network to construct their emulator. Whilst neural networks can be capable tools, they do not provide a complete quantification of the various uncertainties, which can be an essential quality when dealing with complex LSMs.

Attempts to emulate spatio-temporal models without dimension reduction, but by emulating each point in space and time

55   separately, are also popular. Gu and Berger (2016) assume the same correlation function and parameters are shared across all outputs, enabling huge numbers of emulators to be built in parallel. The "emulate every grid cell" approach has been used in environmental applications including aerosol models (Lee et al., 2012; Johnson et al., 2020), tsunami models (Salmanidou et al., 2021) and volcano models (Spiller et al., 2020). These methods ignore spatial and temporal dependence, but rely on speed (though for JULES, the millions of emulators required to follow this approach renders it infeasible). Another approach

60   is to consider space and time as inputs to the model, though this can render the number of data points to large for traditional Gaussian process implementations. Rougier (2008) uses separability in space and time to develop efficient emulators of this type that rely on kronecker identities for their speed.

   Land surface models typically do not exchange information laterally between grid cells (LSMs with more sophisticated hydrology schemes can be an exception to this, where groundwater and rivers can flow between grid cells). Therefore, the

65   structure of the spatio-temporal outputs is often controlled entirely by a set of pre-known forcing data (which could be observational data, or outputs from an atmospheric model, or pre-selected by a practitioner). For example, JULES relies on a set of driving data; providing information about the weather on a sub-daily time step, and various soil properties. In many cases, the land surface model output can be treated as independent in space and time *conditional on the forcing data*. With this framing, the land surface model only outputs a spatial-temporal map because it is input a spatial-temporal map. For example, JULES

70   can be run such that grid cells do not "talk" to each other: what happens in one grid cell has no bearing on what happens in a neighbouring grid cell (except in that their forcing data is likely to be similar, and thus their outputs are likely to be similar).

   In this paper we outline a framework for building emulators of LSMs, leveraging this interesting property of many LSMs to facilitate the emulation of the high-dimensional output. This framework is described in Section 2. We then demonstrate the capabilities of such an emulator by emulating JULES at 1km resolution in Section 3. We use the emulator to obtain a ranking

75   of the relative importance of different model inputs, which can be key to scientific understanding and can guide future model development, and use the emulator to tune the model via History matching. We offer discussion in Section 4.

2   Methods

2.1   Emulators

Gaussian process emulators represent statistical models for functions. The simulator is treated as an uncertain function, and the

80   Gaussian process gives a probability distribution over those functions that can be quickly sampled or evaluated to predict the simulator and to provide uncertainty. A Gaussian process emulator can be written as:

$$y(\theta) \sim GP(m(\theta), k(\theta, \theta)) \tag{1}$$

where y is the output, θ is a vector of input parameters, m is a mean function that can be used to provide prior beliefs about the model, and k is a covariance function. The key feature of a GP is that any finite collection of realisations, $y(\theta_1), \ldots, y(\theta_n)$, is

85   multivariate normal with mean $(m(\theta_1), \ldots, m(\theta_n))$ variance matrix K, where $K_{ij} = k(\theta_i, \theta_j)$. Having observed an ensemble of runs, $\mathbf{y}$, the posterior process, $y(\theta) \mid \mathbf{y}$, is also a GP and, in the standard set up, the mean and covariance function of this posterior GP have a known analytical form (Rasmussen and Williams, 2006). Throughout this article, emulating output y and inputs θ with a Gaussian process will be shortened to: $y(\theta) \sim GP(\theta)$. It can be good practice to re-scale the inputs to all be in $(0,1)$, and to standardise the outputs, and this is also done in this work.

2.2 Emulating a Land Surface Model

We consider the case of an LSM that outputs spatial maps that vary in time. Spatio-temporal maps can be very high dimensional; for example, our study area is Great Britain, where there are 230,615 grid cells (using a grid resolution of 1 km x 1 km). If Gross Primary Productivity (GPP) was output daily then the result would be an approximately 8.5 million dimensional output even if JULES was only run for a single year.

Spatio-temporal correlation structure in LSM's is often inherited through its input forcing data. For example, JULES solves the same differential equations independently in each grid cell and no information is passed horizontally by the solver. As such, we choose to treat an $S \times T$ dimensional output as $S \times T$ different 1D outputs, each with a different set of forcing data inputs (where S represents the number of spatial locations and T the number of timesteps). Mathematically, that is:

$$y_{ts}(W,\theta) = f(W_{ts},\theta), \tag{2}$$

where $y_{ts}(W,\theta)$ is the output at time t and location s, with forcing data W and input parameters θ, f is the land surface model, and W_{ts} is the forcing data only at time t and at location s. Treating the LSM as a model with 1D output, a standard emulator can then be built, that is:

$$f(W_{ts},\theta) \sim GP(W_{ts},\theta). \tag{3}$$

Whilst the assumption of spatial independence (conditional on the forcing data) can be sensible, temporal dependence is often less so. In JULES, various internal state variables are stored and updated at each time step, which provides some temporal structure. For example, the soil moisture (which is modelled by JULES) depends not only on the precipitation at the current time step, but also previous time steps. Another example is the leaf area index: future carbon assimilation and respiration depend on the leaf area index from the previous time step.

Such time structures could be emulated using a 'dynamic emulator' (Mohammadi et al., 2019), but the computational cost of this is prohibitive for the lengths of time LSMs often deal with. In this work, for JULES, we mitigate this time-dependence by working with 8-day averages, which removes the short term temporal dependence. To control for some longer-term temporal dependencies, we supplement the forcing data W_{ts} with a pseudo-forcing variable that records the day of the year. This acts as a proxy for any long-term (i.e. seasonal) temporal structure which our assumption of time-independence ignores.

Any residual temporal correlation is then modelled as residual noise; and so the emulator is no longer forced to perfectly interpolate each simulation. This additional noise is sometimes called a 'nugget', and including it has been found to provide improved emulator performance regardless of whether it is technically needed (Gramacy and Lee, 2012).

2.3 Data abundance

The formulation above in Section 2.2, where each grid cell and each time step is treated as an independent (conditional on the forcing data) data point, results in tens of millions of data points per year for even a single simulation. Theoretically, an abundance of data should greatly improve predictive capabilities. However, Gaussian processes are not designed for large data

sets, as they require the inversion of an $n \times n$ covariance matrix, where n is the number of data points; and so computational time scales with n^3. In computer experiments, roughly 10 data points per input dimension is normally expected and recommended (Loeppky et al., 2009). We could use other statistical models here instead (such as a standard linear regression model, or a neural network), but the flexibility and the uncertainty estimates produced by a GP are desirable.

125   The concept of sparse Gaussian processes was developed to mitigate the computational issues of a GP (Snelson and Ghahramani, 2006). In effect, the idea is to learn some hypothetical, smaller, ideal data set, and then use that to build the emulator. If X is a set of inputs (the combined set of W_{ts} and θ) and $\mathbf{y}$ is a set of outputs, both N in number, then the objective is to learn some set of inputs Z and some set of outputs $\mathbf{u}$ which are only M in number (where $M << N$) whilst maintaining as much accuracy as possible. The smaller set of 'data points' Z and u are called "inducing points". Learning these inducing

130   points, along with all other parameters in the GP, can be done efficiently via variational inference (Titsias, 2009; Hensman et al., 2013, 2015), which involves finding and then optimising a lower bound for the likelihood of the data. The lower bound also facilitates additional speed up by allowing random (small) batches of the total dataset (minibatches) to be used during each iteration of the inference. We make use of the python package `GPflow` (version 1.5.1) (Matthews et al., 2017) which provides an implementation for sparse Gaussian processes. For the GPflow settings, we used 10 inducing points per input dimension

135   and a minibatch size of 1000. Further details for sparse GPs are provided in the appendix, and in the given references.

2.4   Obtaining an Ensemble

To build an emulator of the type described above, we need several simulations of the land surface model. Standard practice would be to run the land surface model for the entire area of interest (for example, Great Britain), and then repeat this several times with different parameter values to obtain an initial ensemble (Murphy et al. (2004); Booth et al. (2012); Williamson

140   et al. (2017)). Our assumption of LSM grid cell independence means that we do not need to run the LSM for the entire area of interest. Instead, we select a subset of coordinates which, collectively, are representative of the different regimes of forcing data that can appear. Each coordinate can then be paired with a different parameter setting.

To see the potential benefits of this approach, consider an example where the area of interest contains 200,000 grid cells, and we can afford to run the LSM for this area 10 times, each time with a different θ value. Alternatively, we could run 2,000,000

145   different θ values, each one run at only one coordinate. Collectively, this should still provide a good coverage of the different possible forcing data, but we also now have *much* better coverage in θ, whilst not requiring any additional simulation effort.

2.5   Emulating JULES

To demonstrate the above framework, we build an emulator for JULES. We narrow our focus into only investigating Gross Primary Productivity (GPP), which is a measure of plant photosynthesis. We begin with GPP as this is the entry point of

150   carbon into the terrestrial carbon cycle. Further work could focus on emulating other aspects of the land carbon cycle (such as net primary productivity, or vegetation and soil stores). The carbon cycle is relevant for studies of climate change, and for these applications JULES is typically run globally with a course spatial resolution (at least 0.5° x 0.5°). However our study was

motivated by an increasing need for detailed process models to inform decisions about land use and management at a much higher resolution. Therefore we run and emulate JULES at a 1km x 1km resolution for Great Britain.

155   JULES grid cells are subdivided into tiles representing vegetated and non-vegetated surfaces. On vegetated tiles, JULES calculates the GPP for different plant functional types (PFTs). The grid cell GPP is a weighted average of the PFT-dependent GPP (depending on the fractional area covered by each PFT). The 5 PFTs we use are: deciduous broadleaf trees (BT), evergreen needleleaf trees (NT), C3 grasses (C3g), shrubs (SH) and cropland (Cr). The fractional area covered by each PFT is set based on land cover data (see Appendix B for details on the JULES simulations). We build 5 independent emulators, one for each

160   PFT. These independent emulators can then be summed together, weighted by the PFT fractions, to provide the final emulator for overall GPP. In summary, the overall emulator is then:

$$\text{GPP}(W_{ts}, \theta) = \sum_{j \in \{PFT\}} \gamma_j \text{GPP}_j(W_{ts}, \theta_j);$$ (4)

with

$$\text{GPP}_j(W_{ts}, \theta_j) \sim GP(W_{ts}, \theta_j),$$ (5)

165   where γ_j is the fraction of PFT j in the grid cell, GPP_j is the PFT-j-specific GPP value, and θ_j is the collection of input parameters that are relevant for PFT j. This overall emulator is simply the sum of 5 distinct emulators, one per PFT. Building one emulator for each PFT and then summing them together makes use of more information known to JULES, which should improve the accuracy of the overall emulator. It also reduces the dimension of θ in each individual emulator.

The PFTs are characterised by parameters describing their physiology, radiative properties, seasonal responses and other

170   attributes. In this configuration of JULES, there are approximately 50 relevant parameters for each PFT. We investigate only 13, 10 of which are different for each PFT, chosen based on previous sensitivity studies of JULES (Booth et al. (2012); Raoult et al. (2016)), and experience working with JULES. Table 1 lists the different tuning parameters θ, standard values, the ranges we consider, and their role in JULES.

Table 1: A table showing the parameters, their "standard" values, their initial un-tuned ranges, and a short description. Most parameters were given different ranges for each plant functional type, indicated by the brackets in the parameter name, these are broadleaf trees (BT), needle-leaf trees (NT), C3 grasses (C3g), shrubs (SH), and crops (Cr).

Parameter	Standard	Range	Description
alpha	0.08	$[0.04, 0.12]$	Quantum efficiency of photosynthesis (mol CO_2 (mol PAR photons)$^{-1}$)
knl	0.2	$[0.05, 0.35]$	Rate of decay of N through the canopy
g_leaf_0	0.25	$[0.1, 3]$	Minimum turnover rate for leaves (360 days)$^{-1}$
dqcrit(BT)	0.09	$[0.045, 0.18]$	Critical humidity deficit (kg H_2O per kg air)
dqcrit(NT)	0.06	$[0.03, 0.12]$	Critical humidity deficit (kg H_2O per kg air)
dqcrit(C3g)	0.051	$[0.0255, 0.102]$	Critical humidity deficit (kg H_2O per kg air)
dqcrit(SH)	0.03	$[0.015, 0.06]$	Critical humidity deficit (kg H_2O per kg air)

dqcrit(Cr)	0.075	$[0.0375, 0.15]$	Critical humidity deficit (kg H_2O per kg air)
f0(BT)	0.875	$[0.65, 0.972]$	Ci/Ca when dq = 0
f0(NT)	0.875	$[0.65, 0.972]$	Ci/Ca when dq = 0
f0(C3g)	0.931	$[0.6916, 1.034208]$	Ci/Ca when dq = 0
f0(SH)	0.875	$[0.65, 0.972]$	Ci/Ca when dq = 0
f0(Cr)	0.8	$[0.5942857, 0.8886857]$	Ci/Ca when dq = 0
g_grow(BT)	20	$[10, 40]$	Rate of leaf growth $(360 \text{ days})^{-1}$
g_grow(NT)	15	$[7, 30]$	Rate of leaf growth $(360 \text{ days})^{-1}$
g_grow(C3g)	20	$[10, 40]$	Rate of leaf growth $(360 \text{ days})^{-1}$
g_grow(SH)	30	$[15, 60]$	Rate of leaf growth $(360 \text{ days})^{-1}$
g_grow(Cr)	20	$[10, 40]$	Rate of leaf growth $(360 \text{ days})^{-1}$
lai_max(BT)	7	$[3.5, 10]$	Maximum leaf area index
lai_max(NT)	7	$[3.5, 10]$	Maximum leaf area index
lai_max(C3g)	3	$[1.5, 6]$	Maximum leaf area index
lai_max(SH)	4	$[2, 7]$	Maximum leaf area index
lai_max(Cr)	3	$[1.5, 6]$	Maximum leaf area index
nmass(BT)	0.0257	$[0.0089, 0.0354]$	Top leaf N content (kgN per kgLeaf)
nmass(NT)	0.01091	$[0.00667, 0.02253]$	Top leaf N content (kgN per kgLeaf)
nmass(C3g)	0.02248	$[0.01076, 0.05433]$	Top leaf N content (kgN per kgLeaf)
nmass(SH)	0.0192	$[0.01, 0.0319]$	Top leaf N content (kgN per kgLeaf)
nmass(Cr)	0.0113	$[0.00565, 0.0226]$	Top leaf N content (kgN per kgLeaf)
rootd_ft(BT)	2	$[0.1, 5.33]$	Parameter for decay of root functioning with depth (m)
rootd_ft(NT)	1.8	$[0.1, 4.8]$	Parameter for decay of root functioning with depth (m)
rootd_ft(C3g)	0.5	$[0.1, 1.333]$	Parameter for decay of root functioning with depth (m)
rootd_ft(SH)	0.5	$[0.1, 1.333]$	Parameter for decay of root functioning with depth (m)
rootd_ft(Cr)	0.5	$[0.1, 1.333]$	Parameter for decay of root functioning with depth (m)
tleaf_of(BT)	278.15	$[273, 283]$	Temperature below which leaves are dropped (K)
tleaf_of(NT)	233.15	$[233, 273]$	Temperature below which leaves are dropped (K)
tleaf_of(C3g)	278.15	$[273, 283]$	Temperature below which leaves are dropped (K)
tleaf_of(SH)	278.15	$[273, 283]$	Temperature below which leaves are dropped (K)
tleaf_of(Cr)	278.15	$[273, 283]$	Temperature below which leaves are dropped (K)
tlow(BT)	0	$[-1, 1]$	Lower temperature parameter for photosynthesis (°C)
tlow(NT)	0	$[-1, 1]$	Lower temperature parameter for photosynthesis (°C)
tlow(C3g)	10	$[9, 11]$	Lower temperature parameter for photosynthesis (°C)
tlow(SH)	10	$[9, 11]$	Lower temperature parameter for photosynthesis (°C)
tlow(Cr)	13	$[11, 15]$	Lower temperature parameter for photosynthesis (°C)
tupp(BT)	32	$[22, 36]$	Upper temperature parameter for photosynthesis (°C)
tupp(NT)	32	$[22, 36]$	Upper temperature parameter for photosynthesis (°C)
tupp(C3g)	32	$[22, 36]$	Upper temperature parameter for photosynthesis (°C)
tupp(SH)	40	$[31, 46]$	Upper temperature parameter for photosynthesis (°C)
tupp(Cr)	45	$[28, 51]$	Upper temperature parameter for photosynthesis (°C)
vsl(BT)	32.50	$[6, 150]$	Regression slope between V_{cmax} and N_{area} (μ mol CO_2 gN^{-1} s^{-1})
vsl(NT)	21.46	$[3, 96]$	Regression slope between V_{cmax} and N_{area} (μ mol CO_2 gN^{-1} s^{-1})
vsl(C3g)	48.03	$[7, 322]$	Regression slope between V_{cmax} and N_{area} (μ mol CO_2 gN^{-1} s^{-1})

Table 2. A table of the accuracy rates of each of the 5 pft specific emulators. A perfect emulator would have an accuracy rate of 95.4%; larger values imply under-confidence, smaller values imply over-confidence.

GPP_{BT}	GPP_{NT}	GPP_{C3g}	GPP_{SH}	GPP_{Cr}
94.8%	95.6%	95.6%	95.4%	95.0%

vsl(SH)	32.16	$[15, 119]$	Regression slope between V_{cmax} and N_{area} (μ mol CO_2 gN^{-1} s^{-1})
vsl(Cr)	20.48	$[9, 95]$	Regression slope between V_{cmax} and N_{area} (μ mol CO_2 gN^{-1} s^{-1})

From here, 20,000 combination of parameters are chosen and paired with 20,000 grid cells. Efforts are made to ensure that
175 a good coverage in θ and W_{ts} is obtained from these choices (by maximising a criteria discussed in the appendix). Running
JULES for these settings then provides the data required to train the emulator outlined above. This emulator is then also used
to perform some exploratory model tuning (Section 3.3) using observational data from two MODIS satellites (Running and
Zhao, a, b) (see Appendix C for details).

3 Results

3.1 Emulator Performance

It is an essential part of the process to check the accuracy of an emulator, just as it is with the land surface model itself. When
training the emulator(s), we hold 10% of the data points aside at random. These held-out data points can then be used to test
the accuracy of the emulator(s). For each of the 5 PFT-specific emulators, we obtain emulator predictions for 1000 randomly
chosen points from the held-out testing data set, and obtain the 2 standard deviation intervals. These 2 standard deviation
185 intervals should approximately correspond with the 95.4% certainty interval; and so roughly 95.4% of the held-out data should
lie within the intervals. Table 2 shows that each of the 5 emulators performs well, with accuracy rates between 94.8% and
95.6%.

With this emulator we can then make fast predictions of 8-day average GPP without needing to run JULES again. We can do
this for any time, location and scenario (whether historical or hypothetical), and for any tuning parameter settings (assuming
190 reasonable ranges) rapidly on a personal laptop. As an example, Figure 1 shows two 1km resolution maps of predicted GPP
for Great Britain, for June 1st 2004, obtained from the emulator, using two different tuning parameter settings.

The emulator is given no information about location, and thus the spatial structure in the predictions is inherited entirely from
the forcing data provided. Other maps like these can be produced for different scenarios representing a range in environmental
data, parameter settings, land use or PFT fractions.
195 The ability to emulate a land surface model at high resolution opens many potential avenues of research. For the rest of this
section we explore two such avenues: sensitivity analysis and calibration.
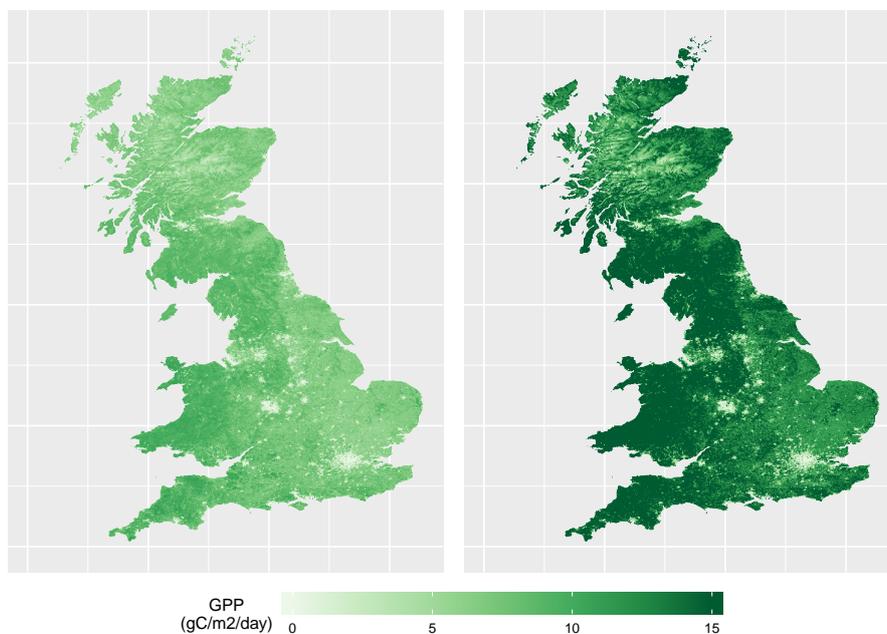
Figure 1. Two maps of GPP for June 1st 2004, as generated by the emulator, for two different parameter settings. The left plot corresponds to a non-implausible set of inputs; and the right plot corresponds to an implausible set of inputs.

3.2 Sensitivity Analysis

With a Gaussian process emulator it is possible to obtain an automatic, preliminary, sensitivity analysis. If the covariance function, k, in Equation (1) is chosen to be a "non-isotropic squared exponential covariance function" (sometimes known as an "automatic relevance determination kernel"); the emulator will automatically obtain estimates of the relative importance of different inputs (i.e. how sensitive the outputs are to individual inputs) if a constant mean function is chosen (Rasmussen and Williams, 2006). This choice of k is available as an option in any Gaussian process software. In training the emulator, lengthscale estimates are obtained, which provide a measure of how far away two points need to be in a given dimension before they become uncorrelated. As such, a smaller lengthscale suggests a stronger relationship between the parameter and the output, and thus greater importance. Figure 2 plots these estimates for each input.

Clearly, each PFT has a different relationship between the inputs and GPP, but some overall patterns are visible. For one, the forcing data is, in general, more important than the parameter settings. The air surface temperature (tas); the gradient of a specific grid cell (slope); the amount of shortwave radiation (rsds); and humidity (huss), all appear important for all PFTs. For tuning parameters, the parameter "tleaf" appears important for broadleaf trees and shrubs, and the parameter "vsl" appears important for all PFTs, as does "alpha" and "nmass". "lai_max" also seems to be important for cropland and C3 grasses. Results like these can provide key insights into a LSM. Whilst more dedicated sensitivity analyses can be conducted, the above can be obtained effortlessly after the construction of a comprehensive emulator like this one.
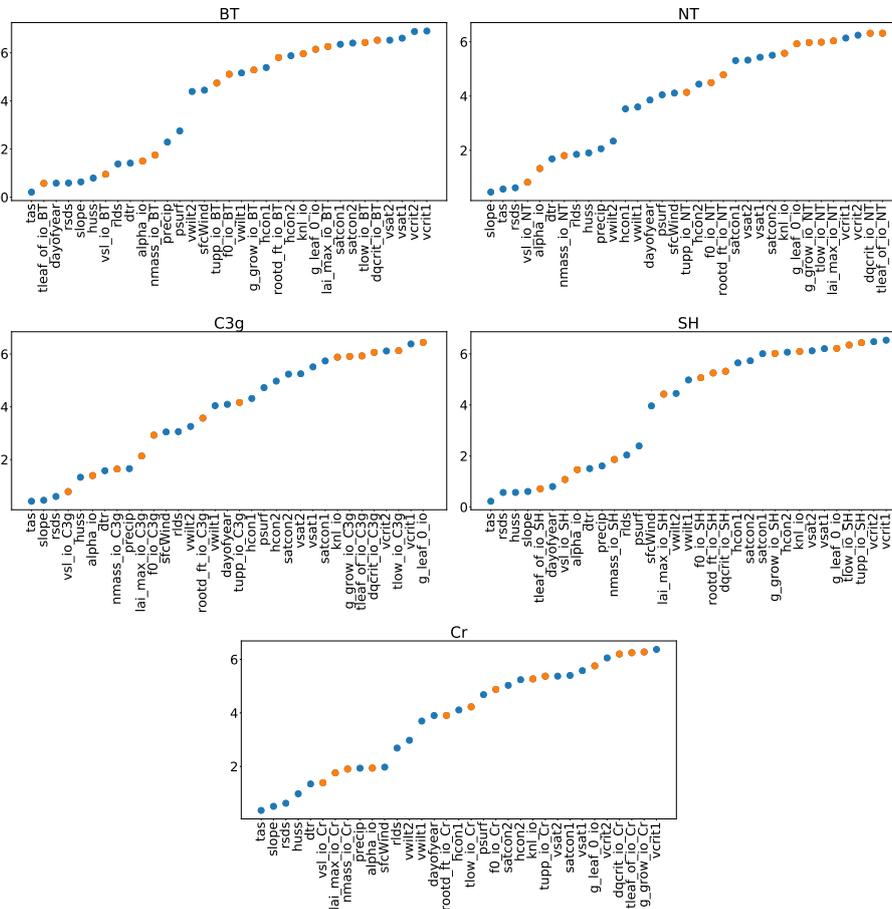
Figure 2. Estimated length scales from the Gaussian process emulators; a smaller value implies a stronger relationship between the input and GPP. Orange points represent tuning parameters and blue dots represent environmental data. Descriptions for the tuning parameter abbreviations can be found in 1, and descriptions for the environmental data can be found in Appendix C (with 1 and 2 representing an upper and lower soil layer).

3.3 Tuning/Calibration

Formal tuning, or calibration, of the various input parameters in a land surface model can be conducted using an emulator.
215 Tuning is the process of choosing the parameters such that the resulting outputs match up with real life observations. Without an emulator, an exhaustive search of the different possible parameter settings is prohibitively expensive, and so tuning in practice involves some degree of arbitrariness, relying heavily on subjective experience and instinct. Alternatively, an optimisation procedure can be taken (Raoult et al., 2016; Peylin et al., 2016), but this can be computationally intensive, the results will not quantify uncertainties completely, and no alternative options are provided if the final result does not agree with scientific belief.

With an emulator, many different parameter settings can be directly tested, facilitating an efficient exploration of the parameter space. History Matching is a straightforward method for testing different parameter settings, ruling out inputs as 'implausible' if observed data does not match with the model output (Craig et al., 1997), and has already been successfully applied to other climate models (Williamson et al., 2013; Couvreux et al., 2021; Hourdin et al., 2021). Given observed data $\mathbf{y}_{obs}$, observational error σ^2_{obs}, a tolerance to model error σ^2_{MD}, a mean prediction for the output of the model $E[\mathbf{y}(\theta)]$ and a predictive covariance of the model output $\Sigma(\theta)$ (the latter two are provided by an emulator); then the implausibility of any given parameter setting θ can be calculated as:

$$\mathcal{I}(\theta) = (\mathbf{y}_{obs} - E[\mathbf{y}(\theta)])^\top \left(\Sigma(\theta) + \sigma^2_{obs}I + \sigma^2_{MD}I \right)^{-1} (\mathbf{y}_{obs} - E[\mathbf{y}(\theta)]), \tag{6}$$

where I is the identity matrix. Implausibility is similar to mean square error, but each grid cell is weighted according to its uncertainty (which has components due to observation error, structural error and emulator variance). A larger implausibility indicates a greater mismatch between the observations and the output from the land surface model (indicating that the parameter setting, θ, that can be ruled out). A conservative threshold for rejecting a parameter setting can be taken as the 99.5% quantile of a χ^2 distribution with l degrees of freedom (where l is the dimension of the observation) (Vernon et al., 2010).

To demonstrate history matching for JULES, we consider a small subset of our grid cells (1000 points chosen to maintain a good coverage of the forcing data), and we randomly sample a time point for each coordinate. We then use the observed GPP data from MODIS (more details in the appendix) for these locations and times, taking the mean of the two observations as the observation. We take the observational error standard deviation, σ_{obs}, as the standard deviation between the two individual observations plus an additional, conservative, 20% of the mean value (because of the inaccuracy of standard deviation estimates obtained from only 2 samples). We obtain a value for the tolerance to model error (model discrepancy) using the three-point Pearson and Tukey formula outlined in Keefer and Bodily (1983) and Revie et al. (2010); using an assumption that real GPP could be up to 40% larger than what JULES outputs, or down to 20% smaller, whilst assuming the output from JULES is still the best guess for what the real GPP could be. These assumptions impose a mild bias (as the hypothetical distribution for the failings of JULES is skewed), which needs to be added to the $(\mathbf{y}_{obs} - E[\mathbf{y}(\theta))$ terms in Equation (6).

We then consider 100,000 different candidate parameter settings (selected using a maximin latin hypercube to ensure a good variety (McKay et al., 1979)); and immediately rule-out roughly 85.5% as implausible, leaving 14475 non-implausible parameter settings.

Some parameter settings didn't get ruled-out because the uncertainty from the emulator is too large. To rectify this, more JULES runs can be made for parameter settings which have not yet been ruled-out, improving the emulator in the non-implausible regions of parameter space, and the tuning process can then be repeated. This is common practice with History Matching, and is called "iterative refocussing" (Hourdin et al., 2021). We obtain another 'wave' of 3000 JULES runs; using 3000 different non-implausible parameter settings (matched with 3000 different coordinates - 2000 chosen as in Section 3.1 and the 1000 coordinates with observations).
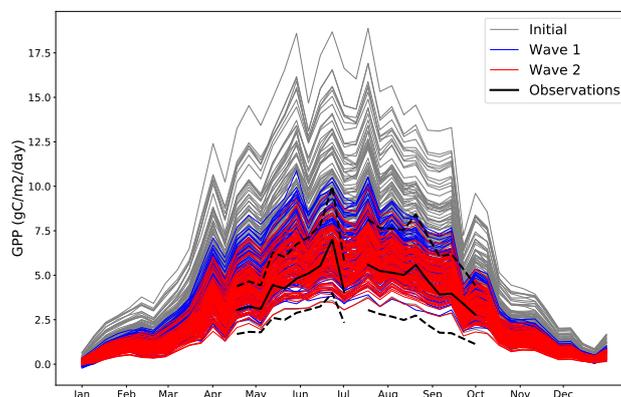
Figure 3. Different time series of GPP for the year 2002 for a random coordinate where we have observations. Grey lines represent 100 randomly chosen parameter settings before any tuning is performed. Blue lines represent 100 randomly chosen parameter settings after the first wave of tuning. Red lines represent 100 randomly chosen parameter settings after the second wave of tuning. The black solid line represents the observations for that coordinate (and the black dashed lines represent the 2 standard deviation intervals of the observational error).

The combined set of the new JULES runs and the previous runs for parameter settings which were not yet ruled-out makes up the second wave's dataset. We fit a new emulator using the same procedure as before, which is then used to further rule-out parameter settings. We find a further 69.9% of parameter settings to be implausible; leaving a total of 4363 non-implausibile parameter settings, or 4.4% of the originally considered set of parameter settings. This procedure could be repeated further, until no changes are found. At any stage, the non-implausible parameter settings represent the parameter choices which sufficiently agree with the observed data and simulations.

Interestingly, we found 'canonical JULES' (i.e., the set of standard values considered) to be non-implausible in the first wave, but with more simulations, we found it be implausible in the second. This suggests that improvements in the standard settings can be made.

The map on the left in Figure 1 was produced using a non-implausible parameter setting, and the map on the right used an implausible parameter setting. There is a clear difference between the two maps. Spatial gradients in the not ruled-out yet parameter setting are much more gentle; with no large extremes in GPP, but still a distinct spatial pattern (for example, lower GPP in the Scottish highlands, and a decrease in GPP going from west to east and corresponding to rainfall gradients). The GPP map corresponding to the ruled-out parameter setting appears very extreme; predicting larger changes over relatively small areas, as well as generally very large values for GPP, with the extremes $> 15gC/m^2/day$.

As an illustrative example of the iterative tuning process; Figure 3 presents different possible outputs from different possible parameter settings.

This plot shows how the distribution of possible trajectories of the model output change after different waves of tuning. By
wave 2, the trajectories closely follow the observations, with their allowed spread appearing to be determined entirely by the
observational error.

4   Discussion

We have outlined a framework for emulating land surface models using sparse Gaussian processes. This framework takes into
account a unique feature of many land surface models, incorporating the information contained within the set of external forcing
data. Under this framework, a substantially better coverage of the input parameters can be obtained in the initial simulation
ensembles, without additional simulation effort and without compromising analytical capabilities. The use of sparse GPs for
emulating a land surface model is the first we are aware of and is natural here where the ensembles are made artificially
large, relative to standard computer experiments. The resulting land surface emulators are incredibly flexible, facilitating fast
prediction, parameter tuning and sensitivity analysis, among other experimental goals. Specific modifications were made to
build an emulator for JULES, but the overall procedure can remain the same for many LSMs. Every land surface model is
different however, and so various modifications to the procedure could be made, depending on the different priorities and
interests of the modeller.

An important experimental goal that could be facilitated with a land surface emulator is that of optimisation; where one
tries to maximise (or minimise) some output. A high resolution land surface emulator would be incredibly useful in informing
decision making on a local level, and this would be a promising research direction. Optimisation using Gaussian processes is
referred to as "Bayesian Optimisation" and already has a thriving literature (Mockus, 2012; Jones et al., 1998). Alternative
procedures which also require emulators to make optimal decisions also exist, inspired by the History Matching methodology
(Lawson et al., 2016).

For the outlined emulator framework, the modification wherein the high dimensional spatial-temporal output is converted to
a large set of 1D outputs allows the information in the forcing data to be readily incorporated and the dimension of the output to
be shrunk. However, removing the time structure (instead of only the spatial) is, in many cases, a simplification rather than an
assumption. Re-adding the time structure may be an interesting direction for future work. This could be via a dynamic emulator
as discussed previously, by using dimension reduction techniques over the time dimension, or by including previous time steps
of the driving data $W_{(t-i)s}$ as inputs to the emulator. The current modifications explained in Section 2.2 seem to provide an
acceptably accurate emulator. Although the 95% certainty intervals produced by the PFT-specific emulators for JULES contain
the truth 95% of the time, the emulators are considerably erroneous on very rare occasions. Also, even though GPP is always
greater than 0, sometimes the emulators will predict GPP to be less than 0 (which is easily rectified by converting any negative
predictions to 0, but this does distort the uncertainty intervals when GPP is predicted small/negative). As such, it should also
be noted that, just like a land surface model, an emulator can always be improved further.

Regarding JULES itself, the preliminary sensitivity analysis in Section 3.2 identified that, in general, the forcing data has
a greater influence on GPP than the parameter settings. This is perhaps obvious, as the amount of plant activity depends

heavily on the environment; but this does suggest improving the accuracy of environmental data should be a key priority for practitioners working with JULES. This agrees with the results from McNeall et al. (2020). The sensitivity analysis performed in Section 3.2 was only a preliminary, automatic, sensitivity analysis. A more comprehensive sensitivity analysis using this emulator (perhaps in the format of Oakley and O'Hagan (2004)) is left for future work.

For the tuning procedure, more waves of simulations could be added until no greater improvements could be made. Alternatively, one non-implausible parameter setting could be hand-selected via expert opinion. Additionally, the observational data used was of relatively poor quality, with a large observational error that hindered the rejection of some poor parameter choices. All low values of GPP (<2 gC/m2/day) were excluded from the observational data because of a mismatch between the two sources of data (Appendix C), which means that many parameter choices which poorly recreate low GPP may not have been rejected. Better observational data would the improve the tuning; but the procedure itself does appear capable.

Similarly, while the above tuning for JULES ensured that the remaining parameter settings match well with observed GPP, this does not necessarily mean that the parameters will match well with other outputs of JULES. Emulating and matching multiple distinct outputs is an interesting avenue for future experimentation with JULES, especially when it is entirely possible that certain parameter settings will be good for one output, but mutually exclusive parameter settings will be good for a different output (this effect is observed by McNeall et al. (2020)). This would mean that no parameter settings will provide an overall good match (the so-called terminal case (Salter et al., 2019)), and suggests the degree of model discrepancy is larger than initially thought. Discovering which observations a land surface model can and cannot reproduce, and which parameter settings are better and which are worse, can be useful information for quantifying and rectifying the flaws in a land surface model.

As an additional note regarding tuning: reducing the set of non-implausible parameter settings does not necessarily impose any individual bounds on the various parameters. For example, if there were only 2 parameters, both promoting plant growth, then it is entirely reasonable to believe a large value for the first and a small value for the second could match with reality, while believing that a small value in the first and a large in the second could also match with reality. What would be important here, is to rule-out small values for both and large values for both. With 53 parameters, rather than just 2, this situation is essentially guaranteed.

A key benefit of the emulator is its computational speed compared to running the land surface model itself. For comparison, simulations of JULES over Great Britain on a 1.5 km x 1.5 km resolution took approximately 7 hours per decade for the spin up (a warm-up period where detailed outputs are not produced) and approximately 15-25 hours per decade in similar model runs to the ones used in this paper. These were run on 72 processors on the Met Office and NERC Supercomputer Node phase 2 (MONSooN2). The emulator however, on a higher 1km x 1km resolution, for Great Britain, takes roughly 20 seconds on a personal laptop for an 8-day prediction using 1 core. This would correspond to approximately 2.5 hours per decade. There are a few important caveats to this which can make the emulator even faster in practice. Firstly, the emulator predictions are fully parallelizable, and so speed increases are easily achieved through using more cores. Secondly, these emulator prediction times are for a personal laptop, which we would expect to perform worse than a supercomputer (MONSooN2 was listed as the 11th most powerful supercomputer in the world). Thirdly, and perhaps most importantly, the emulator doesn't need spin-up

time, and doesn't need to be run for a whole time-series. This makes it easier to directly obtain the desired information, for the desired times and desired locations; long run periods are not always needed. For example, the tuning performed in Section 3.3 did not need the entirety of Great Britain for the entire historical record; and instead the emulator was able to explore many

340 different values of the tuning parameters for a select few times and locations - something which would not be possible without an emulator.

As a final comment from this work, hopefully we have highlighted the relative accessibility of fairly modern statistical techniques. Gaussian process emulators are reasonably sophisticated, but they are also reasonably intuitive, whilst also being incredibly capable. Proportional to the expertise required to actually develop a complex land surface model, relatively basic

345 expertise is required to extract substantially more value via the use of an emulator.

Appendix A: Appendix - Sparse Gaussian Processes

What follows is a technical overview for sparse Gaussian processes and how to fit them. Full discussion and derivation is left to the cited references.

Consider a mean zero Gaussian process (as used throughout this article)[1]. An alternative way to express the GP formulation

350 for an ensemble of model output, $\boldsymbol{y}$, is

$$\boldsymbol{y} \mid \boldsymbol{f}, \sigma^2 \sim \mathrm{N}(\boldsymbol{f}, \sigma^2 I) \tag{A1}$$

$$\boldsymbol{f} \sim \mathrm{N}(\boldsymbol{0}, K_{nn}) \tag{A2}$$

where K_{nn} is the matrix $k(X, X)$. This formulation says that the model data is an observation of an underlying GP, $\boldsymbol{f}$, and will allow for computational simplifications.

355 Posterior inference requires the inversion of K_{nn}, which becomes impractical if n is large, and is particularly problematic for either fully Bayesian inference (which requires this inversion for every proposed parameter sample) but also for methods where the parameters are optimised and plugged in.

A solution to this lies in a variational approximation to posterior inference based on augmenting $\boldsymbol{f}$ with $m << n$ pseudo-simulations, $\boldsymbol{u}$, at input locations Z, called "inducing points". Variational methods propose an approximate parametric posterior

360 distribution, $q(\cdot)$, known as the variational posterior, and find the parameters for $q()$ which minimise the Kullback-Leibler divergence between $q()$ and the true posterior. This is equivalent to maximising a well known lower bound on the marginal likelihood (Salimbeni and Deisenroth, 2017):

$$L = \mathrm{E}_{q(\boldsymbol{f}, \boldsymbol{u})} \left[\log \frac{p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{u})}{q(\boldsymbol{f}, \boldsymbol{u})} \right]. \tag{A3}$$

Taking the variational posterior as $q(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f} \mid \boldsymbol{u}) q(\boldsymbol{u})$, with $q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}; \boldsymbol{m}, \boldsymbol{S})$ (where the slight abuse of notation,

365 $\mathcal{N}(\boldsymbol{u}; \boldsymbol{m}, \boldsymbol{S})$, is used to represent the Normal density for $\boldsymbol{u}$ with mean $\boldsymbol{m}$ and variance $\boldsymbol{S}$) leads to $p(\boldsymbol{f} \mid \boldsymbol{u})$ cancelling in

---

[1]This is a modelling choice that is common with some authors (Binois et al., 2018), whilst more complex mean functions are favoured by others (Bower et al., 2010), with kernel $K(\cdot, \cdot)$. A discussion of these choices can be found in Section 2.7 of Rasmussen and Williams (2006)

Equation (A3), because $p(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{y} \mid \boldsymbol{f})p(\boldsymbol{f} \mid \boldsymbol{u})p(\boldsymbol{u})$:

$$L = \mathrm{E}_{q(\boldsymbol{f}, \boldsymbol{u})} \left[\log \frac{p(\boldsymbol{y} \mid \boldsymbol{f})p(\boldsymbol{u})}{q(\boldsymbol{u})} \right].$$

Additionally, $\boldsymbol{u}$ can be analytically integrated out from $q(\boldsymbol{f}, \boldsymbol{u})$ (required for the expectation) to leave:

$$q(\boldsymbol{f} \mid \boldsymbol{m}, \boldsymbol{S}) = \mathcal{N}(\boldsymbol{f} \mid \mu, \Sigma),$$

370 with

$$\mu = K_{nm} K_{mm}^{-1} \boldsymbol{m}$$
$$\Sigma = K_{nn} - K_{nm} K_{nm}^{-1}(K_{mm} + \boldsymbol{S}) K_{mm}^{-1} K_{nm}^{T},$$

where K_{mm} is the matrix $k(Z, Z)$, and K_{nm} is the matrix $k(X, Z)$. This formulation only requires the inversion of $m \times m$ matrices, providing a significant computational speed-up.

375 Collecting the terms of L that depend on $\boldsymbol{u}$ gives

$$L = \mathrm{E}_{q(\boldsymbol{f} \mid \boldsymbol{m}, \boldsymbol{S})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{f}) \right] + \mathrm{E}_{q(\boldsymbol{u})} \left[\log \frac{p(\boldsymbol{u})}{q(\boldsymbol{u})} \right], \tag{A4}$$

where the second term is the KL divergence between prior $p(u)$ and $q(u)$. The first term can be expressed as a sum of univariate expectations of individual data points so that

$$L = \sum_{i=1}^{N} \mathrm{E}_{q(f_i \mid \boldsymbol{m}, \boldsymbol{S})} \left[\log p(y_i \mid f_i) \right] + \mathrm{E}_{q(\boldsymbol{u})} \left[\log \frac{p(\boldsymbol{u})}{q(\boldsymbol{u})} \right].$$

380 As such, during optimisation of L for the various parameters ($\boldsymbol{m}, \boldsymbol{S}$, Z and the covariance hyperparameters), the data can be sub-sampled at each iteration (still obtaining an unbiased estimator for L). These smaller sub-samples are called minibatches, and provide a second computational speed-up (Hensman et al., 2013).

Implementation for this method is available in the public software `GPflow`.

Appendix B: Appendix - Choosing JULES runs

385 Throughout this article we mentioned how the various JULES simulations were chosen to maintain a good coverage of the forcing data W_{ts} and the tuning parameters θ. The sheer number of runs that were possible under this framework reduces the importance of carefully selected simulations, but we will outline our method here for completeness.

In JULES, because outputs at time t are dependent on state variables from time t-1, a grid cell must be run for the entire time range of interest (rather than only for the time steps required). As such, we could not simply select W_{ts} to have good coverage, 390 we instead had to select a set of grid cells such that the resulting set of W_{ts} had good coverage. To do so we mostly condensed the time series $\mathbf{w}_t$ for a given grid cell s into it's historical mean $\hat{\mathbf{w}}_s$ from 2000-2009, and ensured a good coverage of these means were obtained.

To ensure a good coverage was obtained, we used the combination of three distinct scores. The first score is the minimum distance between two $\hat{\mathbf{w}}_s$. The distance between two $\hat{\mathbf{w}}_s$ represents how similar any two grid cells are, and so maximising this

395 minimum distance M ensures no two chosen grid cells are too similar.

The second score is based on Latin hypercubes. Latin hypercubes are designs where, after binning each dimension d_i, each bin in each dimension is forced to contain only one point (McKay et al., 1979). A Latin hypercube ensures that every dimension has good coverage. As we are not designing our grid cells from scratch, and instead have a larger set to choose from (those that exist in Great Britain), obtaining an actual Latin hypercube is essentially impossible. Instead, we score a potential set of $\hat{W}_s$

400 based on how close to a Latin hypercube it is. Our Latin hypercube score is:

$$L = \sum_{d_i} \sum_{b} |P_b - A_b| \tag{B1}$$

where b is a bin, P_b is the preferred number of points in the bin, and A_b is the actual number of points in the bin. In a real Latin hypercube, there would be as many bins b as desired points in the ensemble, P_b would always equal 1, and A_b would always equal 1. As such, for a real Latin hypercube, $L = 0$. The further away from 0 $\hat{W}_s$ is, the less like a Latin hypercube it is. To

405 make things easier with the number of data points we want, we instead made the number of bins in each dimension equal to 20, and so P_b is equal to $N/20$ (where N is the number of desired ensemble members).

The above two scores together constitute a type of 'maximin Latin hypercube score'. Normally this would be sufficient, but as mentioned above, we have had to work with the historical means $\hat{W}_s$ rather than the actual inputs W_{ts}. The intra-cell variation does still matter for the emulator, as a grid cell which has a very variable climate is more useful - providing more

410 information into how JULES reacts to different forcing values. As such, our third score, C, is is simply the difference between the historical 90th percentile and the historical 10th percentile (summed over each forcing dimension) - providing a measure of how variable that grid cell is.

To select a set of grid cells to run JULES for, we then combine these individual scores as:

$$\lambda_M M - \lambda_L L + \lambda_C C. \tag{B2}$$

415 A set of grid cells with a higher combined score is a better set of grid cells to run. We randomly obtain many potential sets of grid cells and score each of them, the highest scoring set of grid cells is then the one chosen. The λ_i in the above equation are weights to ensure the three component scores are on the same scale, and equal $1/(max(i) - min(i))$.

To choose the set of parameters that go with the grid cells, many Latin hypercubes were generated (which is possible because the tuning parameters θ can take any value in their domain), and the combined design $(\hat{W}_s, \theta)$ with the maximum minimum

420 distance was chosen.

For specifics; more potential designs can be obtained if these schemes are run for a longer period of time, thus providing a better final result. The wave 1 grid cell selection scheme, the wave 1 θ selection scheme and the calibration data grid cell selection scheme were all ran for 5 hours. For the wave 2 design the grid cell selection scheme and the θ selection scheme were each ran for 1 hour. These times were certainly overkill, especially given the large number of JULES runs obtained, and

425 a much shorter run time would have been sufficient.

Appendix C: Appendix - JULES Configuration

The JULES configuration used in this study is JULES version 5.6, based off Blyth et al. (2019). Full details are provided in that paper, but here we provide relevant background information for interpreting the emulator results. We also note any changes from Blyth et al. (2019). The 5 PFTs used were chosen for their relevance to UK landscapes. Three non-vegetated surfaces can also
430 be present in each grid cell: lakes, urban, and bare soil. The fractional coverage of these eight tiles is prescribed in each grid cell based on the CEH Land Cover Map 2000 (Fuller et al. 2002). The soil is represented by four discrete layers (0.0-0.1 m, 0.1-0.35 m, 0.35-1.0 m, and 1.0-3.0 m), with the Van Genuchten approach used for calculating soil hydrology. The relevant parameters are calculated from textures taken from the Harmonised World Soil Database (HWSD: FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012). These parameters are: dry thermal conductivity (hcon), hydraulic conductivity at saturation (satcon), volumetric water
435 content at the wilting point, a critical point (below which GPP is reduced due to soil moisture stress), and the saturation point (vwilt, vcrit, and vsat, respectively). In this work, each layer is treated as a different input to the emulator, and only the top and bottom layers are considered, as the central two mimicked the outer two. Topographic slope is also used to calculate saturation excess runoff; this is derived from a 50 m resolution database - the CEH Institute of Hydrology Digital Terrain Model (IHDTM) (Morris and Flavin, 1990, 1994).

440 The driving meteorology, described in Robinson et al. (2017), is a combination of gridded daily precipitation observations and other meteorological observations. The original data resolution was 40 km, and this was downscaled to 1 km based on topography (Blyth et al. 2019). There are eight meteorological driving variables: the diurnal temperature range (dtr), air temperature (tas) specific humidity (huss), precipitation (precip), air pressure (psurf), longwave radiation (rlds), shortwave radiation (rsds), and wind speed (sfcWind). We take values for these forcing variables from the CHESS data set (Robinson et al.,
445 2017).

The model is run at a half-hour time step. The daily driving meteorology is interpolated to half-hourly using a disaggregation scheme (Williams and Clark, 2014). Model fluxes such as GPP, NPP, respiration, latent heat, sensible heat, and runoff are calculated every half hour. At the end of each day, a phenology scheme updates leaf area index based on temperature functions of leaf growth and senescence. Every ten days, the dynamic vegetation component of JULES (TRIFFID) updates vegetation
450 and soil carbon stores. Competition between vegetation types is turned off in this configuration of the model. Note that in Blyth et al. (2019), LAI was prescribed rather than predicted, and TRIFFID was not used to update carbon stores.

Appendix D: Appendix - Tuning Data

In the tuning procedure, we used the MOD17A2HGF and MYD17A2HGF Version 6 gap-filled GPP data as a test for model tuning and calibration (Running and Zhao, a, b). This is a proof of concept to show the potential for reducing model-data
455 biases through emulators of LSMs. The GPP data is a cumulative 8-day composite of values within 500 m, from the Terra (MOD17A2HGF) and Aqua (MYD17A2HGF) satellites. The data was downloaded from the NASA Application for Extracting and Exploring Analysis Ready Samples (AppEEARS, AppEEARS Team. (2020)). The dataset comes with quality flags, and we only included 'Good quality' data with clear skies and where the main (RT) method was used, yielding either 'best result

possible' or 'good very usable' result. We extracted the 500 m pixel closest to the center of the CHESS grid cell used in the

460   JULES simulation. When both Terra and Aqua had 'good' retrievals for a given time and location, we use an average of the two. An additional constraint was necessary, as sometimes one satellite retrieved a moderately high GPP while the other retrieved a small value (< 2 gC/m2/day). As a conservative fix, we ignored both retrievals when one retrieved a value < 2 gC/m2/day.

Code and data availability. Simulations used to fit the emulator are available at https://doi.org/10.5285/789bea37-0450-4822-9857-3dc848feb937 (Baker et al., 2021). Python code to build and validate the emulator is provided in the supplementary material.

465   *Author contributions.* EB developed the emulator, performed the analyses, and wrote much of the article. AH prepared and ran the simulations, provided insight into JULES, obtained the observational data, and edited the article. DW aided in the development of the emulator, and edited the article. PC aided in the development of the emulator and the general scientific discussion of the problem.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the Natural Environment Research Council [NE/T004177/1].

References

470

AppEEARS Team.: Application for Extracting and Exploring Analysis Ready Samples (AppEEARS), Ver. 2.35., NASA EOSDIS Land Processes DAAC, accessed 2020-02-12 from https://lpdaacsvc.cr.usgs.gov/appeears/, 2020.

Baker, E., Harper, A., Williamson, D., and Challenor, P.: Gross Primary Productivity simulations of Great Britain for emulation, 2001-2010 from JULES land surface model, https://doi.org/10.5285/789bea37-0450-4822-9857-3dc848feb937, 2021.

475 Best, M., Pryor, M., Clark, D., Rooney, G., Essery, R., Ménard, C., Edwards, J., Hendry, M., Porson, A., Gedney, N., et al.: The Joint UK Land Environment Simulator (JULES), model description–Part 1: energy and water fluxes, Geoscientific Model Development, 4, 677–699, 2011.

Binois, M., Gramacy, R. B., and Ludkovski, M.: Practical heteroscedastic gaussian process modeling for large simulation experiments, Journal of Computational and Graphical Statistics, 27, 808–821, 2018.

480 Booth, 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, 2012.

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

Clark, D., Mercado, L., Sitch, S., Jones, C., Gedney, N., Best, M., Pryor, M., Rooney, G., Essery, R., Blyth, E., et al.: The Joint UK Land Environment Simulator (JULES), model description—Part 2: carbon fluxes and vegetation dynamics, Geoscientific Model Development,
485 4, 701–722, 2011.

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., Rio, C., Audouin, O., Salter, J., Bazile, E., Brient, F., Favot, F., Honnert, R., Lefebvre, M.-P., Madeleine, J.-B., Rodier, Q., and Xu, W.: Process-Based Climate Model Development Harnessing Machine Learning: I. A Calibration Tool for Parameterization Improvement, Journal of Advances in Modeling Earth Systems, 13, e2020MS002 217, https://doi.org/https://doi.org/10.1029/2020MS002217, e2020MS002217 2020MS002217, 2021.

490 Craig, P. S., Goldstein, M., Seheult, A. H., and Smith, J. A.: 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, pp. 37–93, Springer, 1997.

Gramacy, R. B. and Lee, H. K.: Cases for the nugget in modeling computer experiments, Statistics and Computing, 22, 713–722, 2012.

Gu, M. and Berger, J. O.: Parallel partial Gaussian process emulation for computer models with massive output, The Annals of Applied Statistics, 10, 1317 – 1347, https://doi.org/10.1214/16-AOAS934, 2016.

495 Hemmings, J., Challenor, P. G., and Yool, A.: Mechanistic site-based emulation of a global ocean biogeochemical model (MEDUSA 1.0) for parametric analysis and calibration: an application of the Marine Model Optimization Testbed (MarMOT 1.1), Geoscientific Model Development, 8, 697–731, 2015.

Hensman, J., Fusi, N., and Lawrence, N. D.: Gaussian Processes for Big Data, in: Uncertainty in Artificial Intelligence, p. 282, Citeseer, 2013.

500 Hensman, J., Matthews, A., and Ghahramani, Z.: Scalable Variational Gaussian Process Classification, in: In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, 2015.

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, 2008.

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., Musat, I., Fairhead, L., Diallo, F. B., and Volodina, V.:
505 Process-based climate model development harnessing machine learning: II. model calibration from single column to global, Journal of

Advances in Modeling Earth Systems, n/a, e2020MS002 225, https://doi.org/https://doi.org/10.1029/2020MS002225, e2020MS002225
2020MS002225, 2021.

Johnson, J. S., Regayre, L. A., Yoshioka, M., Pringle, K. J., Turnock, S. T., Browse, J., Sexton, D. M. H., Rostron, J. W., Schutgens, N.
A. J., Partridge, D. G., Liu, D., Allan, J. D., Coe, H., Ding, A., Cohen, D. D., Atanacio, A., Vakkari, V., Asmi, E., and Carslaw, K. S.:
510   Robust observational constraint of uncertain aerosol processes and emissions in a climate model and the effect on aerosol radiative forcing,
Atmospheric Chemistry and Physics, 20, 9491–9524, https://doi.org/10.5194/acp-20-9491-2020, 2020.

Jones, D. R., Schonlau, M., and Welch, W. J.: Efficient global optimization of expensive black-box functions, Journal of Global optimization,
13, 455–492, 1998.

Keefer, D. L. and Bodily, S. E.: Three-point approximations for continuous random variables, Management Science, 29, 595–609, 1983.

515   Kennedy, M. C. and O'Hagan, A.: Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical
Methodology), 63, 425–464, 2001.

Lawson, A., Goldstein, M., and Dent, C.: Bayesian framework for power network planning under uncertainty, Sustainable Energy, Grids and
Networks, 7, 47–57, 2016.

Lee, L. A., Carslaw, K. S., Pringle, K. J., and Mann, G. W.: Mapping the uncertainty in global CCN using emulation, Atmospheric Chemistry
520   and Physics, 12, 9739–9751, https://doi.org/10.5194/acp-12-9739-2012, 2012.

Loeppky, J. L., Sacks, J., and Welch, W. J.: Choosing the sample size of a computer experiment: A practical guide, Technometrics, 51,
366–376, 2009.

Lu, D. and Ricciuto, D.: Efficient surrogate modeling methods for large-scale Earth system models based on machine-learning techniques,
Geoscientific Model Development, 12, 1791–1791, 2019.

525   Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J.:
GPflow: A Gaussian process library using TensorFlow, Journal of Machine Learning Research, 18, 1–6, http://jmlr.org/papers/v18/16-537.
html, 2017.

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

530   McNeall, D., Williams, J., Betts, R., Booth, B., Challenor, P., Good, P., and Wiltshire, A.: Correcting a bias in a climate model with an
augmented emulator., Geoscientific Model Development, 13, 2487–2487, 2020.

Mockus, J.: Bayesian approach to global optimization: theory and applications, vol. 37, Springer Science & Business Media, 2012.

Mohammadi, H., Challenor, P., and Goodfellow, M.: Emulating dynamic non-linear simulators using Gaussian processes, Computational
Statistics & Data Analysis, 139, 178–196, 2019.

535   Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A.: Quantification of modelling
uncertainties in a large ensemble of climate change simulations, Nature, 430, 768–772, 2004.

Oakley, J. E. and O'Hagan, A.: Probabilistic sensitivity analysis of complex models: a Bayesian approach, Journal of the Royal Statistical
Society: Series B (Statistical Methodology), 66, 751–769, 2004.

Petropoulos, G., Griffiths, H., Carlson, T., Katidis, P. I., and Holt, T.: SimSphere model sensitivity analysis towards establishing its use for
540   deriving key parameters characterising land surface interactions, Geoscientific Model Development, 7, 1873–1887, 2014.

Peylin, P., Bacour, C., MacBean, N., Leonard, S., Rayner, P., Kuppel, S., Koffi, E., Kane, A., Maignan, F., Chevallier, F., Ciais, P., and Prunet,
P.: A new stepwise carbon cycle data assimilation system using multiple data streams to constrain the simulated land surface carbon cycle,
Geoscientific Model Development, 9, 3321–3346, https://doi.org/10.5194/gmd-9-3321-2016, 2016.

Raoult, N. M., Jupp, T. E., Cox, P. M., and Luke, C. M.: Land-surface parameter optimisation using data assimilation techniques: the
545   adJULES system V1.0, Geoscientific Model Development, 9, 2833–2852, https://doi.org/10.5194/gmd-9-2833-2016, 2016.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning., MIT Press Cambridge, MA, 2006.

Revie, M., Bedford, T., and Walls, L.: Evaluation of elicitation methods to quantify Bayes linear models, Proceedings of the Institution of
       Mechanical Engineers, Part O: Journal of Risk and Reliability, 224, 322–332, 2010.

Ritchie, P. D., Harper, A. B., Smith, G. S., Kahana, R., Kendon, E. J., Lewis, H., Fezzi, C., Halleck-Vega, S., Boulton, C. A., Bateman, I. J.,
550   et al.: Large changes in Great Britain's vegetation and agricultural land-use predicted under unmitigated climate change, Environmental
       Research Letters, 14, 114 012, 2019.

Robinson, E., Blyth, E., Clark, D., Comyn-Platt, E., Finch, J., and Rudd, A.: Climate hydrology and ecology research support system
       meteorology dataset for Great Britain (1961-2015) [CHESS-met] v1.2, https://doi.org/10.5285/b745e7b1-626c-4ccc-ac27-56582e77b900,
       2017.

555 Rougier, J.: Efficient emulators for multivariate deterministic functions, Journal of Computational and Graphical Statistics, 17, 827–843,
       2008.

Running, S. W. and Zhao, M.: MOD17A2HGF MODIS/Terra Gross Primary Productivity Gap-Filled 8-Day L4 Global 500 m SIN Grid
       V006, NASA EOSDIS Land Processes DAAC, accessed 2020-02-12 from https://doi.org/10.5067/MODIS/MOD17A2HGF.006., a.

Running, S. W. and Zhao, M.: MYD17A2HGF MODIS/Aqua Gross Primary Productivity Gap-Filled 8-Day L4 Global 500 m SIN Grid
560   V006, NASA EOSDIS Land Processes DAAC, accessed 2020-02-12 from https://doi.org/10.5067/MODIS/MOD17A2HGF.006., b.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P.: Design and analysis of computer experiments, Statistical Science, 4, 409–423, 1989.

Salimbeni, H. and Deisenroth, M.: Doubly stochastic variational inference for deep Gaussian processes, arXiv preprint arXiv:1705.08933,
       2017.

Salmanidou, D. M., Beck, J., and Guillas, S.: Probabilistic, high-resolution tsunami predictions in North Cascadia by exploiting sequential
565   design for efficient emulation, Natural Hazards and Earth System Sciences Discussions, 2021, 1–24, https://doi.org/10.5194/nhess-2021-
       63, 2021.

Salter, J. M., Williamson, D. B., Scinocca, J., and Kharin, V.: Uncertainty quantification for computer models with spatial output using
       calibration-optimal bases, Journal of the American Statistical Association, 114, 1800–1814, 2019.

Sexton, D. M. H., Murphy, J. M., Collins, M., and Webb, M. J.: Multivariate probabilistic projections using imperfect climate models part I:
570   outline of methodology, Climate Dynamics, 38, 2513–2542, https://doi.org/10.1007/s00382-011-1208-9, 2012.

Sexton, D. M. H., Karmalkar, A. V., Murphy, J. M., Williams, K. D., Boutle, I. A., Morcrette, C. J., Stirling, A. J., and Vosper, S. B.: Finding
       plausible and diverse variants of a climate model. Part 1: establishing the relationship between errors at weather and climate time scales,
       Climate Dynamics, 53, 989–1022, https://doi.org/10.1007/s00382-019-04625-3, 2019.

Snelson, E. and Ghahramani, Z.: Sparse Gaussian processes using pseudo-inputs, in: Advances in neural information processing systems,
575   pp. 1257–1264, 2006.

Spiller, E. T., Wolpert, R. L., Ogburn, S. E., Calder, E. S., Berger, J. O., Patra, A. K., and Pitman, E. B.: Volcanic Hazard Assessment for an
       Eruption Hiatus, or Post-eruption Unrest Context: Modeling Continued Dome Collapse Hazards for Soufrière Hills Volcano, Frontiers in
       Earth Science, 8, 396, https://doi.org/10.3389/feart.2020.535567, 2020.

Titsias, M.: Variational learning of inducing variables in sparse Gaussian processes, in: Artificial Intelligence and Statistics, pp. 567–574,
580   2009.

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

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.

585 Williamson, D. B., Blaker, A. T., and Sinha, B.: Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model, Geoscientific Model Development, 10, 1789–1816, 2017.