Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2021-205-RC1, 2021

Comment on gmd-2021-205

Anonymous Referee #1

Referee comment on "Emulation of high-resolution land surface models using sparse Gaussian processes with application to JULES" by Evan Baker et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-205-RC1, 2021

Review of paper: 'Emulation of high-resolution land surface models with sparse Gaussian processes with application to JULES'

General Comments:

This paper develops a framework for statistical emulation of a land surface model (LSM) that can produce predictions with uncertainty at a high resolution and demonstrates the approach and its capabilities through a case study with the Met Office's LSM, JULES. The approach takes advantage of the fact that information is not exchanged laterally between grid-boxes in the model during a simulation. The model's output is treated as independent in space and time (via working with 8-day averages as output), conditional on a set of external forcing data, which includes information about weather-related variables and various soil properties.

Here the authors use the theory of sparse Gaussian Processes, which involves a variational approximation to the posterior inference, to build their emulator, creating an emulator that is fast to use for prediction at high resolution and very flexible. The paper then demonstrates the capabilities of this emulation approach for the JULES model for both sensitivity analysis and parameter tuning via History Matching. The discussion is interesting and covers both the strengths of the approach as well as its limitations.

This is a novel approach to emulation for a land surface model, and has significant potential to improve high resolution model prediction with uncertainty for use in supporting local decision making for policy. The paper definitely falls within the scope of GMD and EGU. It is written to a high standard, and I have enjoyed reading it. However, there are several points that I believe need clarification (see specific comments below) – Once these issues/comments are addressed, then I would recommend the publication of the manuscript in GMD.

Specific Comments:

- Line 8 (in abstract): '…acknowledges the forcing data…' Would a general reader of the abstract know what is meant by the term 'forcing data' here? I think it would be useful to clarify this.

- Line 28-29: '…(with 20 years of spin up simulation… took 1 additional hour)…' This might be a naïve question, but I find this is a little confusing – why does 20 years of spin-up simulation only take one hour, when it takes 30 hours to simulate 11 years of simulation?

- Line 58: '… these methods ignore spatial and temporal dependence, but rely on speed…' I don't think this quite tells the full story for these methods – they do ignore spatial and temporal dependence in the emulator models, but under the assumption that spatial/temporal dependence carries through to a reasonable extent via the training data that is used to construct the emulators. Please edit the sentence here to reflect this.

- Line 89: 'standardise the outputs, and this is also done in this work'. There is no detail on how outputs are standardised in the application later on in the paper? For reproducibility, please clarify how this is done.

- Line 103: Is the vector of forcing variables, w_{ts}, just being treated like extra parameters for the emulator fit here? Please make this clear in the text. [I think it is, but it took me a while to realise this.]

- Line 111: 'working with 8-day averages'. Is the input data for the 'forcing variables' also converted to 8-day averages to match the output y? - So that the temporal nature of both is on the same resolution for the emulator?

- Line 112: '…we supplement the forcing data W_{ts} with a pseudo-forcing variable that records the day of the year…' I don't understand this. How is this done? How does this 'pseudo forcing variable' link to the 8-day average? [Where is this day in the 8-day average? Day 1? Centred (4 or 5)?...] How is it used in the application? (It isn't mentioned in Sect 2.5.) More detail would be useful.

- Section 2.3, Lines 117-121: This paragraph seems to imply (via the use of the term 'data point' in the sentences) that the 'n' data points of the n x n covariance matrix inversion for a Gaussian Process in Section 2.2 (line 121) is of the same magnitude as the number (tens of millions) of data points that a single simulation produces across the full grid (line 119). Is that true? Please clarify the relative magnitude of 'n' for the matrix inversion that is encountered for the application.

- Lines 129 / 134: Are the inducing points Z used for the application a subset of the training data points X from the JULES simulations that are run, or can they be at (W_{ts}, θ) combinations that are outside the training set?

- Line 133: '…during each iteration of the inference.'. It is not clear to me how the inference is actually carried out? Through an MCMC-type procedure? What is involved in an 'iteration of the inference' when fitting the emulator? Please can you supply a little more information to make it easier to understand how the method is actually applied? [What is GPflow doing?]

- Line 143: I think it would be very useful to reference the Appendix B here, and connect your approach to selecting the training data for your application in to this section (2.4).

- Line 154: Can you also state the number of grid-cells in total that cover Great Britain at this resolution here? – to make clearer the full extent of the spatial grid being covered.

Section 2.5: Please can you add more information to this section to explicitly describe what the 'forcing data' for the model actually is? How many variables? What are they? Maybe list them in a table format like Table 1? I think this is important to the application, especially with the sensitivity plots in Section 3.2 showing that many of these

environmental variables are highly important (more important than the uncertain parameters), yet there is no description of it until very briefly when you get to the end of Appendix C, which is a bit late.

- Line 174: '20,000 combinations of parameters are chosen and paired with 20,000 grid cells'. How many training 'data points' does this actually lead to for fitting the emulators? Stating this clearly will give the reader a full appreciation of the amount of training data you have and so a better understanding of quantities such as '10% of the data points' on line 181…

- Line 190-191: '… shows …maps of predicted GPP for Great Britain, for June 1st 2004, obtained from the emulator,…'. How do you get a map of the output (GPP) for a specific day, when the GPP output used for the emulator is an 8-day average? More detail/clarity on how the emulator is sampled to produce the map is needed.

- Line 198-203: Are these the choices of the covariance function k and mean function m for the fitted emulators? It would be very useful for re-producibility if these choices were clearly stated somewhere in the Methods section (Section 2).

- Line 205: Should 'each input' at the end of this sentence actually be 'each PFT'?

- Line 238-243: 'We obtain a value for the tolerance to model error…' I don't understand fully what is done here – This needs more explanation. The formula in the references looks to estimate the 5^{th}, 50^{th} and 95^{th} percentiles of a distribution on the discrepancy term – Are the '20% smaller' and '40% larger' equated to those 5^{th} and 95^{th} percentile points? What is used for the 50^{th} percentile point? - Is that coming from the model (JULES, or the emulator?) or the MODIS observations? Is a different value of the tolerance to model error (and so different 'mild bias') obtained for each observed grid-cell? How is the actual value of the 'mild bias' obtained?

- Line 244: '100,000 different candidate parameter settings'. I think that these are only sampled over θ? And that the forcing variables are fixed for the selected s,t observation cases? – Is that correct?

- Line 250: How are the 3000 non-implausible parameter settings to be run through JULES in the next wave selected from the 14475 that you have? Is there a way to ensure these are well-spaced over the non-implausible parameter space? Or, do you just target those with the highest emulator uncertainty? Or choose them randomly? This isn't clear.

- Line 250: '2000 chosen as in Section 3.1' I cannot see in Section 3.1 that co-ordinates are matched to parameter settings. Should this be 'as in Appendix C'?

- Line 266: '…very large values for GPP values, with the extremes >15gC/m^2/day…' This should be '8-day averaged GPP', yes? Check here and throughout the the other sections that the output is described correctly. i.e. the raw daily GPP and the 8-day average GPP is not the same.

- Line 340 (with Line 233-235): '…explore many different values of the tuning parameters for a select few times and locations…'. How robust are the results in Section 3.3 to the selection of the observational data used for the comparison? Has this been tested?

- Line 372-373: 'The formulation only requires the inversion of m x m matrices…'. In the equation on line 372, the first matrix inversion is for 'K_{nm}'. I think the subscript notation here needs to be amended to 'mm': so, 'K_{mm}'. Also, I may be wrong, but this formulation looks to be following that in Equation 8 of Salimbeni and Deisenroth (2017) – please check the sign of the '**S**' within the brackets in the equation on line 372, as in the Salimbeni

paper Eq 8 this has opposite sign.

- Line 390: What is meant by 'mostly condensed'?

- Line 393: 'The first score is the minimum distance between two w_s hat vectors'. Is this the minimum distance between any two w_s hat vectors from a selected potential set of these vectors? Please make this clearer in the text.

- Line 410-412: 'C is simply the difference…'. Is this calculated for each individual grid cell? If so, is it then summed over the grid cells in each potential set, and then summed again over the different forcing dimensions to produce a single number C for that set? Also, when summing over the forcing dimensions, are the values for each dimension scaled in some way so that a forcing quantity on a scale with larger magnitude isn't overweighted in the sum? Please clarify.

- Line 418-419: What is the dimension of θ for these Latin hypercubes over θ? Please add this to the text. [I think it is 53 (all PFT parameters are treated separately here, and then the design will collapse down when fitting the emulators for each individual PFT, using just the PFT's subset of θ?)]

Technical Corrections:

- Line 60: Change 'to large' to 'too large'

- Line 79-80: The sentence: 'The simulator is treated as…' needs clarity, as at the start it says 'an uncertain function' for the simulator, but then this is referred to as 'those functions' on the next line in the same sentence.

- Line 135: Change '…in the appendix,… ' to '…in Appendix A,…'

- Line 159: 'See Appendix B for…' Should this be Appendix C?

- Tables at the top of Page 8: This is a formatting thing, but it looks weird for the end of Table 1 to come after Table 2 here? Please put Table 2 after Table 1.

- Line 174: Change 'combination' to 'combination**s**'

- Line 175: Change '…in the appendix' to '…in Appendix B,…' [Make it clear which section the reader should go to for the information.]

- Line 176: Remove the word 'also'.

- Line 178: '(see Appendix C for details).' Should this be Appendix D?

- Line 200: Replace ';' with ','

- Figure 2 caption, Line 3: Change: '…can be found in 1, and…' to '…can be found in **Table 1**, and…'

- Line 231: Change: '…, θ, **that** can be ruled out.' to: '…, θ, can be ruled out.'

- Line 242: There is a missing square bracket on the expectation of y term in the formula.

- Line 269: No need for a new paragraph here?

- Line 310: Should 'Appendix C' be changed to 'Appendix D'?

- Line 410: '…third score, C, is is simply…'. Remove the second 'is'.

- Remove the second 'appendix' from each of the Appendix sections titles. It isn't needed.