Response to Reviewers

***CEC1**:*

Dear authors,
in my role as Executive editor of GMD, I would like to bring to your attention our Editorial version 1.2:
https://www.geosci-model-dev.net/12/2215/2019/
This highlights some requirements of papers published in GMD, which is also available on the GMD website in the 'Manuscript Types' section:
http://www.geoscientific-model-development.net/submission/manuscript_types.html
In particular, please note that for your paper, the following requirements have not been met in the Discussions paper:

- "Code must be published on a persistent public archive with a unique identifier for the exact model version described in the paper or uploaded to the supplement, unless this is impossible for reasons beyond the control of authors. All papers must include a section, at the end of the paper, entitled "Code availability". Here, either instructions for obtaining the code, or the reasons why the code is not available should be clearly stated. It is preferred for the code to be uploaded as a supplement or to be made available at a data repository with an associated DOI (digital object identifier) for the exact model version described in the paper. Alternatively, for established models, there may be an existing means of accessing the code through a particular system. In this case, there must exist a means of permanently accessing the precise model version described in the paper. In some cases, authors may prefer to put models on their own website, or to act as a point of contact for obtaining the code. Given the impermanence of websites and email addresses, this is not encouraged, and authors should consider improving the availability with a more permanent arrangement. Making code available through personal websites or via email contact to the authors is not sufficient. After the paper is accepted the model archive should be updated to include a link to the GMD paper."

Therefore please provide a the source code which exactly relates to the version used for this publication in a permanent archive (DOI). Additionally, you should provide the data (training data + output data). If the amount of the data is too high, please state so in the data availability section and provide the information on which data has been used (similar as in the github repository) within the data availability section.
Yours,
Astrid Kerkweg (GMD executive Editor)

Public Answer:

Dear Astrid Kerkweg,

Many thanks for clarifying the guidelines of code and data storage requirements for our development submission. To complement the GitHub repository for full reproducibility, code, models and details on the experimental configuration for reproducing results of

the three machine learning emulators are now stored in a permanent and public OSF repository (DOI: 10.17605/OSF.IO/8567D). The DOI will be added to the data availability statement. We have further requested DOIs for all training and testing data sets that will be published on the ECMWF server and ready for download on request. As soon as we have received the DOIs, we will add them to the data availability statement.

We hope this addresses your concerns and we will gladly take additional steps otherwise.

Sincerely
Marieke Wesselkamp (for all authors)

Final Answer:

We thank the editor again for supporting our submission procedure with this comment. We have received the DOIs for the now publicly available training and test data sets and added them to the data availability statements.

RC1:

Referee Review of 'Advances in Land Surface Model-based Forecasting: A Comparison of LSTM, Gradient Boosting, and Feedforward Neural Networks as Prognostic State Emulators in a Case Study with ECLand'

The authors have identified a component of numerical land surface and weather forecasting that has not previously been tested against current methods of surrogate model development. It is the role of this paper to develop (and provide links to the code of) surrogate models and verify them against a benchmark numerical model.

There is a lot to like about the paper, of course the points below focus on weaknesses, but I would like to thank the authors for an enjoyable read, and some very good research.

Overall I think the points below constitute minor revisions (or appropriate rebuttals from the authors), but I cannot recommend the paper for publication as is.

Related work and appropriate references are included for machine learning methods. Though there are very few references to examples of the numerical experiments on land surface that the surrogate models can provide for.

In terms of scientific quality and significance, the paper expertly develops relevant machine learning methods for predicting variables of land surface models, which is an important and challenging step toward a complete evaluation of the surrogate models.

The full significance of the paper is currently understated because the authors do not provide examples of how the surrogate models could be applied to numerical experiments on land surfaces, and, importantly, how the inaccuracies quantified

through comparison with ECLand could impact such experiments. I recommend the authors revise the paper so that such examples and related discussion are included – this could be in the discussion section. In addition, I have made further recommendations below.

Public Answer:

Dear Simon O'Meara,

Many thanks for assessing our work and providing many valuable suggestions for its improvement. The general comment gives a very helpful perspective, sorry for not having pointed out the significance of surrogates to the full extent. We will address the comment together with the last specific comment and the comment of referee 2.

In our revised manuscript, we will place the development of our emulators more clearly in the context of coupled earth system models: In the IFS, the land surface is coupled to the atmosphere via skin temperature, the predictability of which is known to be influenced by soil moisture and soil temperature. This is the numerical interface where a surrogate model could act in application and it motivates the experiment from a broader perspective, within which we also mention their application as adjoint models. Currently however, only a subset of ECLand variables is represented by the emulators so they don't replace the full numerical model capabilities.
As such, we will continue to point out that the emulators are useful as alone standing models for the aforementioned experiments on the land surface. The computation of forecast horizons is an example in this context, as we can see it as a step toward a seasonal predictability analysis of land surface components. A full predictability analysis requires ensemble simulations, and the emulators can serve here again as a quick surrogate for the numerical model (will be added as example). We will also mention sensitivity analysis in an uncoupled version in this context (will be added as example).
Alongside this however, we will address the last specific comment and therefore stress that before we can use the emulators for any such experiments as a reliable alternative, an evaluation on observations is necessary to avoid misleading statements. We will underline this point by referring to two specific sources of error in a basic emulation procedure: That is the structural uncertainty by statistical approximation of the numerical model, and the training and inference in the currently synthetic data domain.

We hope this will address some of your concerns.

Kind regards
Marieke Wesselkamp (for all authors)

Final Answer:

As anticipated, we have now addressed the content related concerns like described in the public answer. We added a section 4.5 to the discussion, i.e. Experimentation with Emulators, where we target these concerns together with the concerns of Referee 2.

Lines 110-113
It is unclear whether the surrogate models developed here can predict all of the variables that the original ECLand model predicts (and could therefore potentially fully replace the ECLand model).
Final Answer:

We have added the information right after the description of variables in line 113.

I do not see in the main paper information on the runtime of surrogate models (for experiments representative of numerical experiments on land surfaces) alongside the runtime for ECLand for comparison. This information does need to be included as it is the driving force behind the work.

Public Answer:

Will be included in the document.

Final Answer:

This information is now included at the beginning of the results section (section 3), with approximate evaluation runtimes for the three emulators and for ECLand. We refer to this section in the introduction to discussion, when we discuss the significance of runtime improvements.

As mentioned by the other reviewer, although a link to GitHub is provided, a persistent public archive source is not provided.

Public Answer:

See answer to editorial comment.

Final Answer:

Sic.

Citations in the main text are very messy – a mixture of citation styles, making it unacceptable for publication in its current form.

There are multiple spelling and punctuation errors that need resolving before publication.

Public Answer:

We will of course clean the citations and the spelling errors.

Final Answer:

Citations and hopefully spelling mistakes are all corrected and in the same style now.

The abstract describes the emulators as reliable alternatives, however, the discussion stresses that the definition of reliability depends on the application (thereby placing the determination of reliability on the reader). As such, I recommend the abstract be changed to accurately represent this important discussion point.

Public Answer:

See answer to general comment. We will adjust our abstract to better match the revised content.

Final Answer:

We thank the reviewer again for noting this inconsistency. We have addressed this concern in the abstract by stating that reliability depends on the application of emulators.

Where necessary '-3' to denote per unit cubed needs to be superscript

Public Answer:

Will be adjusted.

Final Answer:

We excuse for this flaw; all physical units have been adjusted now.

In figure 2 and elsewhere, the type of fraction that snow cover fraction represents needs to be stated, e.g. (%) or (0-1)

Public Answer:

Will be adjusted.

Final Answer:

Adjusted in Figure 2 and 3, as well as in variable descriptions in methods and in results.

Section 3 and throughout – RMSEs and MAEs should be given in units of the variable they are assessing model accuracy for, e.g. K for soil temperature.

Public Answer:

Will be adjusted.

Final Answer:

Adjusted: In the table stated in captions and in the text units are added where necessary.

Because RMSE and MAE have units of the variable they are assessing model accuracy for, I do not think that RMSE and MAE results of different variables can be combined, as I think they are in Figure 2a and Table 2 and in other parts of results (e.g. Fig. 4a). The main text should be changed accordingly.

Public Answer:

We thank the referee for making this point, and we agree that the aggregated RMSE and MAE scores are not meaningful for inference. However, as we conduct a multi-objective and unweighted optimization towards the global average during model training with the MSE, the aggregated results we report also indicate the global test scores. We state in the discussion that the results on single variables may even differ with a variable-targeted optimization. As such, we prefer to keep reporting the global aggregated scores but will point out their lack of interpretation in the discussion.

Final Answer:

Sic.

Because ACC is a relative value I can see how ACC results of the assessed variables can be combined into one score per model. If this combination is what is shown in Figure 2 (and perhaps elsewhere) then it needs to be stated clearly. Additionally, it should be explained in the method how ACC results of the difference variables were combined, e.g., is an arithmetic mean calculated?

Public Answer:

We thank the referee for this observation. The ACC is calculated as the spatial arithmetic mean over grid cells for the forecast horizon, and as the spatio-temporal mean for the total scores we report. We will add the description of aggregation formally in the methods section.

Final Answer:

As we noted that we stated this at the beginning of the methods section for the other scores already, we simply added this as another sentence after the definition of the ACC.

The caption in figure 3 needs to explain what the top row of sub-plots is showing, i.e. average snow cover in these regions – but what kind of average and from what source is the data, is it ECLand?

Public Answer:

Will be adjusted.

Final Answer:

We added the information on the top row of the subplots to the figure caption.

There needs to be greater emphasis in the abstract and elsewhere that when accuracy is discussed, the authors mean in terms of verification against synthetic data, not evaluation against observations. I think the authors should state very clearly somewhere that further work is needed for evaluation against observations before recommendation of any of the surrogate models for numerical experiments is possible.

Final Answer:

We thank the author very much for his detailed and helpful assessment. We noted that the statement on synthetic data is already in the introduction and, as anticipated in the answer to general comment, we highlight the uncertainties arising with this to the new discussion section, Experimentation with Emulators.

We thank the reviewer again for this helpful assessment. See the answer to general comment.

Simon O'Meara

RC2:

General comments
This paper describes a comparative analysis of emulators as surrogate models for land surface modeling. All three tested emulators achieved high predictive scores. Different effectiveness and the unique advantages of each emulator are analyzed and discussed. This presented work shows the great potential of emulators in land surface modeling, especially regarding computational effectiveness. The authors did a great job in describing the models and in explaining the training and testing procedures. The logic of this paper is quite clear, and it is very well written. I only have a few very minor points for the authors to consider.

(very) Minor
I know the emulators are tested as offline surrogate models, but some discussions on the potential use of the emulators within the fully coupled model could guide the usage of the emulators in future research and development.

Public Answer:

We thank the referee for the generous assessment and this comment on our work. We acknowledge it and will address this as described in the answer to general comment.

Final Answer:

We thank the referee again for the comments. We addressed the content-related concerns together with the concerns of Referee 1 with a new discussions section 4.5 on Experimentation with Emulators.

Technical corrections
L50 and elsewhere: please update the reference format.

Public Answer:

Will be adjusted.

Final Answer:

Adjusted.

Table 1: how do you feed 'low' and 'high' into the emulators?

Public Answer:

Grid cells are dived into multiple fractions of the different coverage types, of which high and low vegetation without snow each are one. So, they are given to the emulators as percentage values between 0 and 1.

Final Answer:

Sic.

I think Section 2.3.2 is a nice and concise description and summary of LSTM.

Final Answer:

We thank the reviewer for the generous assessment!