Point-by-point response to referee comments on revised submission

Quantifying uncertainty in simulations of the West African Monsoon with the use of surrogate models

Matthias Fischer, Peter Knippertz, Roderick van der Linden, Alexander Lemburg, Gregor Pante, Carsten Proppe, John H. Marsham

We would like to thank the reviewer for the final remarks and we want to comment on those. Technical remarks are corrected.

- Section 2.1.1, Para1: Lines 203-205: 'Since probability varies strongly across the input space, it is meaningful to train the model with higher accuracy in regions with higher probability. This is because we construct surrogate models particularly for performing global sensitivity analysis.'
- Lines 206-207: 'Thus, for a more accurate sensitivity analysis, it is crucial for the model to exhibit higher accuracy in regions of the parameter space where the PDF values are greater.'

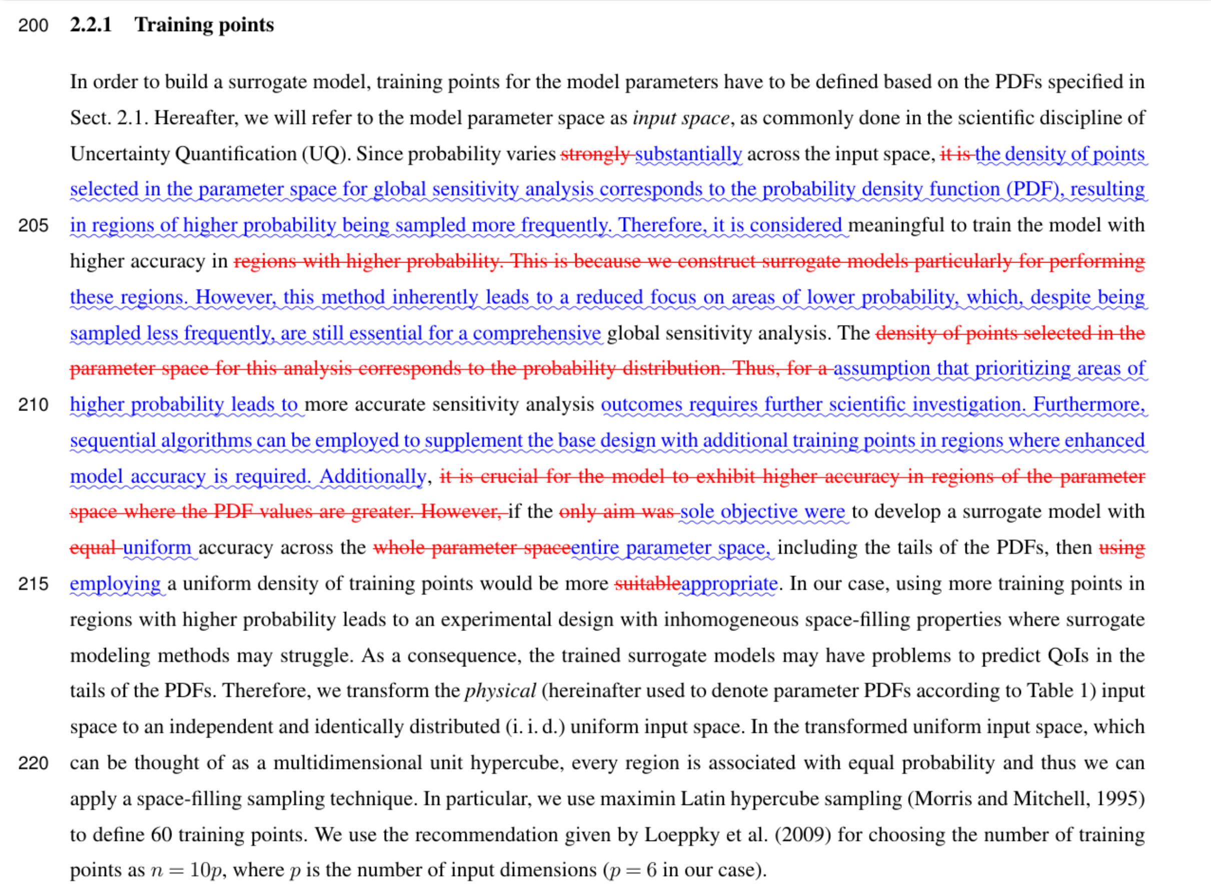
I think these statements are in some ways misleading and this paragraph should be edited to remove any misconceptions / ensure clarity and to also acknowledge the potential negative consequences of the sampling strategy applied. Following the author's response to my previous comment on the sampling for the training data (on training the model with higher accuracy in regions with higher probability, previously lines 195-196), I'm still not convinced about the validity of this. I can understand that the authors want the surrogate to be as accurate as possible where they will sample more, but with a fixed amount of training data in total (as I think is the case here – using n = 10*p (where p is the number of parameters perturbed), this approach must have an opposite effect on the accuracy of the samples in areas of lower probability (by reducing it), which although sampled less, **can/will still be sampled in a global sensitivity analysis and so can affect the results of it**. There is **no evidence** provided (or to my knowledge) to say that this approach / sampling strategy is **crucial** to obtain a more accurate sensitivity analysis, and I think it is just as possible that it could lead to less accurate sensitivity results.
The fact that the authors intend to perform a global sensitivity analysis doesn't make sense to me as a reason to vary the accuracy of the underlying model (here, the emulator/surrogate model) that you want to understand the sensitivity of. The effects of the PDFs are still accounted for in the sampling of the sensitivity analysis procedure itself, and so this seems to be an unnecessary step that has potential to induce possibly significant inaccuracy in some emulator predictions and hence the obtained sensitivity results. When constructing a surrogate model, technical aspects such as changes in the smoothness of the surface that one is trying to approximate can affect the emulators accuracy around the input space and so be valid reasons for the requirement of more/less training data in different areas – In my experience, if more data is needed, this is added in addition to the base training sample of size 10*p. Given this, it seems also possible that the outcome of the sampling strategy described could lead to fewer training points in areas of input space that the Gaussian process might already find the output more difficult to capture well [if they happen to be the areas of lower probability], which would then further lead to poor representation of the climate model, which could affect sensitivity results.
I understand that it isn't possible (due to computational expense) to re-run the study with a uniform training sample for the surrogate model and do the direct comparison, and also that validation of the surrogate models should provide some evidence that the emulator prediction is reasonable across input space [this evidence seems limited here in showing prediction accuracy in different areas of input space]. However, I think it is important that any caveats of the sampling strategy used are

clearly acknowledged [i.e. that the global sensitivity analysis **can/will** still sample in areas of low probability, where the emulator here will be less accurate, which could adversely affect the resulting sensitivities] and that all statements of something being ‘better’ or ‘crucial’ are either evidenced or not used.

Thank you for the very detailed comment. We have clarified the limitations of this methodology in the manuscript so that incorrect conclusions are avoided. In particular, the argument that an experimental design with a density of training points equal to the probability density function (i.e. higher density in more probable regions) would be optimal for a global sensitivity analysis (GSA) has been weakened as this requires further methodological investigations that cannot be carried out at this point. We highlighted that regions with a sparse distribution of training points (i.e. the tails of the PDFs) can strongly influence the outcome of a global sensitivity analysis, even though they are sampled less frequently.

Furthermore, we added the possibility to add further training points if needed by sequential sampling techniques depending on the model accuracy.

Finally, we emphasize that the explained methodical step should be skipped, if a surrogate model with equal probability within predefined input parameter ranges was desired. In that case a uniform density of training points (e.g. a standard Latin hypercube design) may be used.



2.2.1 Training points

In order to build a surrogate model, training points for the model parameters have to be defined based on the PDFs specified in Sect. 2.1. Hereafter, we will refer to the model parameter space as *input space*, as commonly done in the scientific discipline of Uncertainty Quantification (UQ). Since probability varies ~~strongly~~ substantially across the input space, ~~it is~~ the density of points selected in the parameter space for global sensitivity analysis corresponds to the probability density function (PDF), resulting in regions of higher probability being sampled more frequently. Therefore, it is considered meaningful to train the model with higher accuracy in ~~regions with higher probability. This is because we construct surrogate models particularly for performing~~ these regions. However, this method inherently leads to a reduced focus on areas of lower probability, which, despite being sampled less frequently, are still essential for a comprehensive global sensitivity analysis. The ~~density of points selected in the parameter space for this analysis corresponds to the probability distribution. Thus, for a~~ assumption that prioritizing areas of higher probability leads to more accurate sensitivity analysis outcomes requires further scientific investigation. Furthermore, sequential algorithms can be employed to supplement the base design with additional training points in regions where enhanced model accuracy is required. Additionally, ~~it is crucial for the model to exhibit higher accuracy in regions of the parameter space where the PDF values are greater. However,~~ if the ~~only aim was~~ sole objective were to develop a surrogate model with ~~equal~~ uniform accuracy across the ~~whole parameter space~~entire parameter space, including the tails of the PDFs, then ~~using~~ employing a uniform density of training points would be more ~~suitable~~appropriate. In our case, using more training points in regions with higher probability leads to an experimental design with inhomogeneous space-filling properties where surrogate modeling methods may struggle. As a consequence, the trained surrogate models may have problems to predict QoIs in the tails of the PDFs. Therefore, we transform the *physical* (hereinafter used to denote parameter PDFs according to Table 1) input space to an independent and identically distributed (i. i. d.) uniform input space. In the transformed uniform input space, which can be thought of as a multidimensional unit hypercube, every region is associated with equal probability and thus we can apply a space-filling sampling technique. In particular, we use maximin Latin hypercube sampling (Morris and Mitchell, 1995) to define 60 training points. We use the recommendation given by Loeppky et al. (2009) for choosing the number of training points as $n = 10p$, where p is the number of input dimensions ($p = 6$ in our case).