Review Report on "A Gaussian process emulator for simulating ice sheet-climate interactions on a multi-million year timescale: CLISEMv1.0"

The manuscript describes a coupled emulation approach called CLISEMv1.0, which builds two separate emulators for an ice sheet model (AISMPALEO) and a climate model (HadSM3) – The outputs from these two emulators provide inputs to each other, enabling synchronous simulation of ice sheet evolution and climate changes. The authors conducted several sensitivity analyses regarding how the two emulators are built (e.g. depending on how the ice sheet input for the climate emulator is defined), how long the coupling time is, and how the lapse rate is adjusted to account for the elevation difference between the climate model grid and the ice model grid. While the coupled emulation approach itself is scientifically highly important and perhaps long overdue, the current coupled emulation results shown in Sections 3 and 4 need a lot of further clarification before the manuscript can be considered for being published in GMD. More specific comments are listed below.

Major comments:

1. The description about the three different experimental design setups (EMULATOR_70, EMULATOR_100a, and EMULATOR_100b) in Section 2.3 and the result of coupled experiments described in Section 3.1 are contradicting to each other. According to Section 2.3, "EMULATOR_70 has a good spread between the different ice volumes and ice areas" but somehow the results described in Section 3.1 and also shown in Figure 9 indicate that the design EMULATOR_70 seems to have some serious flaw. The other two are described as designs with some notable flaws in Section 2.3, but somehow lead to better results. The manuscript gives some brief description on this issue in Section 3.1, but the authors did not really get to the bottom of the issue – In fact I cannot find any good rationale for how the ice model settings for EMULATOR_100a, and EMULATOR_100b are determined at the beginning – Why did the author decided to let EMULATOR_100a have "more small ice sheet geometries (ice volumes) compared to EMULATOR_100b (Figure 4) and has a good spread for the ice area of the input ice sheet geometries" and EMULATOR_100b be "poorly defined by ice sheet area as there are several experiments with the same ice sheet area yet different ice sheet geometry, but is well defined for ice sheet volume"? Are they some data of opportunity from some other experiments? Or did the author gradually add more model runs to these to designs until they give some sensible results shown in Figure 9? I think the authors need to describe their decision making process behind these design points in detail.

2. Related to the above point, it is hard for me to understand why the coupled emulation based on EMULATOR_70 leads to such poor results. For example, why does it lead to ice volume change that is largely unresponsive to the CO_2 concentration change when ice volume is used? Similarly, to me it is hard to figure out the true reason for the poor results in Figure 9c for all three design schemes. If an emulator based on two parameters (ice volume and area in this case) leads to a worse result than an emulator based on only one parameter for the *same perturbed physics ensemble*, the only possible explanation is that the emulator with the two parameters failed to capture the behavior of the original simulator. I suspect the poor results stem from the poor emulation accuracy when the emulators are built on both ice volume and ice area. In fact, Table 1a shows that there are only **11** design points for the two ice parameters (ice area and ice volume) and, to make the problem worse, these two parameters are highly correlated. I think any emulation approaches are destined to fail with such a small number of design points that are highly correlated with each other.

3. I am not sure why Section 4 is called ‘Bayesian’ sensitivity analysis because nothing in the section seems to be particularly ‘Bayesian’. There seems to be no consideration on uncertainty in the model parameters in the form of posterior densities, which is typically done in Bayesian analysis. In addition, the authors somehow decided to throw away the emulators and build a new time series model for uncertainty quantification. Is there any particular reason behind this decision? I am also wondering if there is any particular reason that only the first-order autocorrelation is considered here.

Minor Comments:

1. Lines 204-215: Related to the Major Comment 1 above, I think describing EMULATOR_100a and EMUATOR_100b as ‘bad’ emulator seems to be weird. I think this part can be improved by clarifying how EMULATOR_100a and EMUATOR_100b are actually designed; otherwise, readers may wonder why the authors decided to use ‘bad designs’. Later, in Section 3, they will be surprised by the fact that these ‘bad designs’ led to better results.

2. Lines 239-240 “Therefore, they might be doing a poor job in reconstructing the simulated temperatures well.”: I think ‘well’ should be deleted.

3. Line 259-260: “The notion of ice sheet parameter as an emulator input is introduced in previous studies to be an integer ranging from 1 to the number of ice sheet geometries”: The sentence does not make much sense. Please revise.

4. I feel that the overall writing quality of Sections 3 and 4 are notably worse than that of the other Sections. Hopefully the authors can improve the texts in the revised version.