First, I apologize to the authors for taking so long to complete this review of “PALEO-PGEM v1.0: A statistical emulator of Pliocene-Pleistocene climate” for publication in GMD. In this paper, the authors describe the building of a statistical emulator from the PLISM-GENIE intermediate complexity model in order to provide a continuous reconstruction of important climate variables over the last 5 millions years, using interpolation to downscale these variables to a higher resolution. This paper logically follows a number of previous articles describing both building of emulators of intermediate complexity models (e.g., Holden et al. 2014) and their application to a range of important scientific questions, such as the links between climate and biodiversity (e.g. Range et al. 2018). It adds in particular a representation of ocean dynamics, which was lacking in previous studies, via the use of PLASIM-GENIE, the coupling of the two models having been described previously (Holden et al. 2018). Overall, the paper is well written and the methodology is sound. I thus recommend publication provided that some important points are taken care of or better discussed:

1. There is no evaluation of the impact of the addition of ocean dynamics on the climate reconstructions, despite being one of the main justifications of the paper (l. 48-49, l. 430). How different would be the results using PLASIM-ENTS rather than PLASIM-GENIE? This leads me to another question: in the PLASIM-GENIE experiments with varying ice sheets, is the land/sea mask changed accordingly or are only the height/albedo feedbacks taken into account (with no varying coastlines impacting ocean dynamics)? If the land/sea mask is modified, I wonder how valid are the transition between ocean modes linked to varying sea level in the subsequent emulations of the climate?

2. Related to the above, the anomaly method for the downscaling approach supposes that the biases between an observed climate state Ct and the emulated climate Et remain somehow constant through time. This is a very crude assumption given, among others, the documented changes in ocean circulation dynamics across the last 5 Ma. How well is the ocean circulation represented in the model? And ocean circulation changes, e.g. between the LGM and MH?

3. I am a bit disappointed by the lack of actual comparison to data over the last 5 Ma. For instance, how does it represent the mid-Pliocene period (and compare to the extensive database of PlioMiP exercises) or glacial-interglacial T and P variations? And transient events like MIS M2?

4. I think a concise but very general reminder of what is an emulator and of the theoretical basis behind could be useful for non specialized readers. In the present version, this reduces to one sentence (l. 58-59) because the Sections 3 and 4 either use too many reference to previous papers (Section 3) or are probably too specialized already (Section 4).
5. The discussion on the limits of the approach and of the (numerous) uncertainties and approximations made should be expanded, in particular if the aim is to provide a widely available climatic reconstruction that is then used to force ecological niche modelling or biodiversity models (l. 36-42). Because this requires at least some confidence in the ability of the model to reproduce “true” (absolute) paleoclimatic conditions and variability (see points 1-3 above).

Other minor details l. 52 “naïve simulation would not be possible for an application of this ambition”. I do not understand this sentence.

l. 272-274. “GPs are highly flexible non-parametric regression models which have greater modelling power than linear models”. Please clarify.

l. 393. “Warm biases are more modest”. On the basis of the figure, this is hard to believe as there are regions with a warm bias as large as the cold bias of other regions.

Fig. 3 and 4. Please use the same color scale and range as the PMIP ensemble to ease comparisons. Why is the Southern Ocean cooling rather than warming in the MH simulation?