Interactive comment on **“PALEO-PGEM v1.0: A statistical emulator of Pliocene-Pleistocene climate”** ***by*** **Philip B. Holden et al.**
Crucifix (Referee) michel.crucifix@uclouvain.be

We thank Michel Crucifix for this thorough and useful review. Our responses are in bold face and the associated revisions to the manuscript are in italics.

The authors propose a latitude-longitude reconstruction of the climate of the whole Pleistocene, using a Gaussian process emulator calibrated on two experiment designs with the PLASIM-GENIE model. It uses use CO2 and sea-level as inputs, based on an inverse modelling reconstruction provided by Stap et al. and, where available, ice core observations. R code with input files are provided.

The process for designing and calibrating the emulator is largely based on earlier work (experiment design, PCA emulator). There is however a cunning novelty: using two similar experiment designs for isolating the climate anomaly caused by ice sheets.

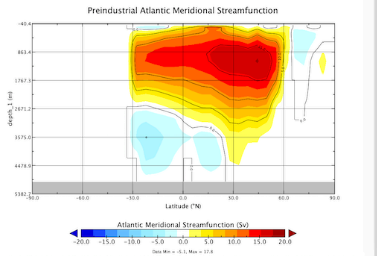
Although perhaps not in line with the reviewer ‘etiquette’, I wish to make a personal comment, in the hope that the editor and authors will be forgiving find this intrusion useful for the evaluation of the work under concern here. After the articles of Araya-Melo et al. (2015), Bounceur et al. (2015), and Lord et al. (2017), we had everything in place at UCLouvain to provide a similar reconstruction, and in fact we tried a few. What stopped us from publishing are:

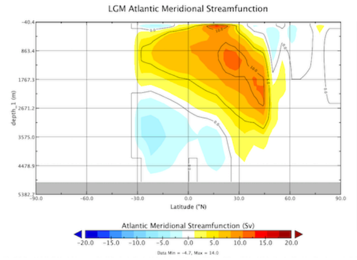
- that the ocean circulation in LOVECLIM was not behaving adequately, with stronger, deep ocean circulation at glacial maxima, at odds with proxies for ocean ventilation. At some stage we thought of mending the simulation with an additional freshwater perturbation, but this work was never finalised to the point of publishing.

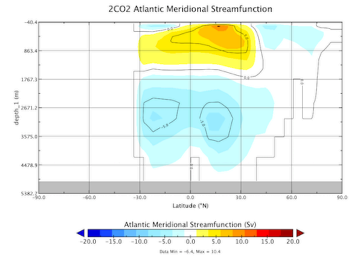
  We have added a paragraph to discuss the LGM and 2xCO2 AMOC:

  The climate sensitivity of the optimised parameter set is 3.2°C. The maximum Atlantic overturning is 17.8Sv, at a depth of 1.1km with the 10Sv contour, an indicator of the location of NADW formation, at a latitude of 56°N. Under LGM forcing, Atlantic overturning weakens to a peak of 11.1Sv at a depth of 1.0km and the 10Sv contour shifts southward to 45°N. Under doubled CO_2 forcing, Atlantic overturning weakens substantially to a peak of 7.6Sv at a depth of 0.4km.

  The Atlantic meridional streamfunctions are shown below for your interest - we have not included these plots in the revised manuscript, but would be happy to do so if the reviewers or editor would like.







- the strategy used in Araya-Melo et al. 2015 of summarising ice sheet forcing with a single quantity, and which is applied here by Holden et al., can be problematic. There is nothing to guarantee that Weschelian ice sheets were located similarly to those of the Early Pliocene, and also, as the authors rightly acknowledged, the build-up and decay phases of ice sheets are quite asymmetrical during the late Pleistocene. This might not be that much of a problem in certain applications, but it can be heavily misleading to users who would use this product in Europe or in Siberia without much discernment. For example, I would be particularly worried of archaeologists using the provided reconstructions near the limits of ice margins. A 3-D reconstruction of the Pleistocene can be very popular, so it needs to be disseminated wisely.

  We agree and have now included an extensive Section 9 detailing the assumptions and weaknesses of the approach (see below)

This little experience brings me to the following, and related comments, about the present contribution by Holden et al.

1. It definitely needs to come up with appropriate health warnings about usage limits of the reconstruction. This is particularly crucial since the introduction presents it

This is a very good point, and we have now included a section detailing the assumptions and weaknesses of the approach. We feel strongly that this approach should be published for dissemination, and would welcome alternative model variants (such as derived from LOVECLIM) to better quantify modelling uncertainties. Ecologists have great need of this data and they have long used (or "misused") whatever paleo-climate estimates are available. As you note yourself, such data are in very high demand. However, spatially-explicit time series of models are rarely available, usually only a few time-slices, so that a common practice is to apply linear interpolations in space and time. Our estimates are therefore a substantial improvement from the best available existing approaches (at least prior to 140ka). Macroecologists are interested in broad-scale spatiotemporal patterns, usually using >1x1 degree cells, and >1Ky time interval and, in general, the dynamics matter much more to them than the exact temperature/precipitation values. We expect that ecologists will understand very well that our paleo-climate data are estimates, following the general principle of modelling (all are wrong, but some are useful), although we certainly agree with the reviewer that detailing the weaknesses of the approach will be of substantial benefit as ecologists may not appreciate the specifics, and have added the new Section 9:

9 Limitations of the approach

PALEO-PGEM is to our knowledge the first attempt to provide a detailed spatiotemporal description of the climate of the entire Pliocene-Pleistocene period. It is essential to understand the main limitations of our modelling framework, discussed below, some of which may induce large errors or uncertainties in specific applications, or even rule out certain applications completely. For all practical purposes and for the foreseeable future, substantial uncertainties exist in any paleoclimate reconstruction as a result of incomplete knowledge, computing limitations and irreducible climatic noise. Ideally, these uncertainties *should be quantified in relation to any reconstruction and their implications propagated through the analysis. Our approach provides an estimate of inherent uncertainty derived from the emulation step of the reconstruction and thus underestimates the full uncertainty, but nevertheless in some aspects remains comparable to the uncertainty in state-of-the-art reconstructions of particular periods as measured by the variance across ensembles of PMIP simulations.*

Compared to state-of-the-art models, PLASIM-GENIE is a relatively low resolution, intermediate complexity climate model. This implies that processes operating at spatial and temporal scales below the native resolution of the climate model cannot be properly represented, although certain aspects of spatial variation are reintroduced in a highly idealised way by the downscaling process. The temporal effects of dynamical processes operating at sub-millenial timescales are further filtered out by the approximation inherent in the emulator construction that the climate is in quasi-equilibrium with the forcing, which is then only resolved at 1000-year time intervals.

In applications where (downscaled) time-slice simulations are adequate and are available from higher complexity models and/or multi-model ensembles (Section 7), these would normally be preferable to PALEO-PGEM as errors and biases will generally be smaller, particularly in high latitudes, regions of steep topography, close to coastlines or in known regions of locally extreme climate. We note that HadCM3 climate simulations (Singarayer et al 2017), downscaled to 1° resolution are available back to 120 kaBP (Saupe et al 2019), which would provide preferable (or supplementary) climate data for applications restricted to this time-domain.

The emulator uncertainty captures much of the uncertainty seen in multi-model intercomparisons (Figures 3 and 4), but PALEO-PGEM cannot fully represent model uncertainty, because it is derived from a single configuration of a single model. Most clearly in this respect, the 90% uncertainty range of climate sensitivity (3.8 ± 0.6°C) is understated relative to multi-model estimates of 3.2 ± 1.3°C (Flato et al 2013). Some significant biases in spatial patterns are also apparent, most clearly temperature biases in high southern latitudes.

Emulator forcing is limited to orbit, CO_2 and ice sheets. Ice meltwater forcing is not considered so that millennial variability, especially important in North Atlantic, is neglected. The land-sea mask and orography are held fixed, so that ocean circulation changes driven by changing gateways (e.g. the closing Panama isthmus, with implications for the thermohaline circulation) are neglected and feedbacks driven by changing orography are neglected, especially important in regions of rapid tectonic uplift.

The representation of ice sheets applies Peltier 5G deglaciation ice sheets (Peltier 2004), assuming a fixed relationship between global sealevel reconstructions (derived from benthic oxygen isotopes) and the spatial form and extent of ice sheets. This approximation neglects the substantial asymmetry between build-up and decay phases of ice sheets and assumes that ice sheets were located similarly in all previous Pliocene-Pleistocene glaciations, which may not have been the case. Particular caution is therefore essential when applying the climate reconstruction at locations near to the margins of ice sheets.

We apply a downscaling approach because spatial climate gradients can be critically important for ecosystem dynamics, especially in mountainous regions which are poorly resolved at native climate model resolution (Rangel et al 2018). The downscaling approximation assumes that the lapse rate within a downscaled grid cell does not change with time, but it does capture the first order effect of topographic complexity by assuming a constant present-day lapse rate. Similarly, the downscaling cannot capture feedbacks between atmospheric circulation and high resolution topography, which could alter the patterns of rain shadowing. However, for many applications, it is preferable to neglect this second order feedback than to neglect the first order effect of a rain shadow that could not be resolved at native climate model resolution (e.g. the Atacama), which downscaling imposes through the baseline climatology. Other simplifications include the implicit assumptions of fixed mountain glaciers and ecotone distributions. In short, the high-resolution reconstructions should not be interpreted as a faithful reconstruction of high-resolution climate, but serve to introduce a more realistic degree of spatial variability.

2. a product oriented to end-users including archaeologists and biodiversity experts. A critical evaluation of the validity of the reconstructions near the North Atlantic (with emphasis on ocean circulation effects) needs to be provided.

See new section 9 above

In the introduction it is clearly said that uncertainty attached to the emulator (here, as a surrogate of GENIE-PLASIM) is distinct from the model uncertainty. This is true, and hence what would think that the evaluation (or "validation") of the emulator (as a surrogate) should be clearly distinguished from the evaluation of the model as a representation of the real climate. I found that this important distinction is pretty blurred in the section 6 (strangely divided into a section heading an a subsection 6.1). In fact there is very little about the evaluation of the emulator as a surrogate of GENIE-PLASIM. The authors refer to the PMIP ensemble and feel comfortable that the emulator-based reconstruction is in broad agreement with PMIP simulation of the mid-Holocene and the LGM, but in doing this the authors are mainly evaluating the reconstruction, not the emulator as a statistical surrogate. And, as I suggest in point 1. above, this evaluation is not providing with non-climatologist users with enough information about its application domain (the dos and don'ts). It is also quite uncomfortable that the emulator provides so- called bioclimatic variables (MIN and MAX over the seasonal cycle) while the "validation" is made on the basis of seasonal averages.

Our main interest here is to derive useful reconstructions for paleo-applications, and so we regard a comparison of the emulated outputs with existing reconstructions as the most important test, capturing simultaneously the climate model errors and the emulation errors. We think that the faithfulness of the emulator wrt the simulator will be of less direct interest to ecologists and other users, but have expanded the cross-validation section in order to better quantify sources of emulation error and have added analysis to cross-validate the seasonal and annual average emulators used in the comparisons with multi-model intercomparisons:

Table 3 summarises the cross-validation of the eight emulators (i.e. four bioclimatic variables, two forcing categories). The second column tabulates the percentage of variance explained by the leading ten principal components, $\sum_{c=1,10} V_c$, and represents the maximum variance that could be explained by the emulators if they were perfect. The remaining columns tabulate the metric P when building the emulator with a series of different covariance functions, being the alternatives available in the DiceKriging R package (Roustant et al 2012). The reduction in variance explained (relative to column 2) reflects additional errors due to emulation.

The temperature decompositions explain 94-99% of the ensemble variance, compared to 87-90% for the precipitation decompositions. Under emulation, the variance explained is 81-98% for the temperature fields and 73-83% for precipitation fields. The emulator performance is weaker for precipitation, because the low order components needed to explain much of the ensemble variability are more difficult to emulate.

The power exponential was found to give comparable or better performance compared to the other covariance functions in all eight emulators and was therefore chosen as the default covariance function, and used in all analysis that follows.

Table 4 summarises the variance explained under cross-validation of the seasonal and annual average emulators used in the following Section 7. DJF (JJA) temperature emulator performance is similar to Min (Max) temperature emulator performance, suggesting that northern hemisphere temperature is more difficult to emulate than southern hemisphere temperature, as would be expected for the ice-sheet emulator in particular. The performance of the various seasonal precipitation emulators is similar (82.7% to 84.8% for the orbit and CO_2 emulator, 72.4% to 75.4% for the ice-sheet emulator), but annual precipitation is easier to emulate than seasonal precipitation (88.6% for the orbit and CO_2 emulator, 81.9% for the ice-sheet emulator).

	DJF	*JJA*	*Max*	*Min*	*Mean*
Orbit and CO_2 emulator					
Precipitation	*84.8%*	*83.9%*	*82.7%*	*82.7%*	*88.6%*
SAT	*95.0%*	*97.8%*	*98.1%*	*95.0%*	*96.7%*
Ice-sheet emulator					
Precipitation	*74.0%*	*72.4%*	*75.4%*	*73.3%*	*81.9%*
SAT	*82.1%*	*94.8%*	*95.1%*	*80.9%*	*90.4%*

Table 4. Seasonal and annual mean emulator performance (as used in Section 7), measured by the metric P (Eq. 3, including ten components). A power exponential covariance is used in all cases. Note that max and min values repeat data from Table 3.

3. Downscaling. Is this correct that downscaling as presented here assumes con- stant sub-grid correction anomaly, defined as the difference between the present- day observations and simulated grid-box-mean in a reference experiment (the anomaly being on the log of precipitations in the low-precipitation areas)? This treatment is arguably inadequate in palaeoclimate applications, where topogra- phy, surface type (think of Swiss glaciers to take

but one example), ice sheet mar- gins, land-sea mask, and ecotone boundaries vary substantially. Again, aren't we misleading the users by providing the illusion of a high-res reconstruction, while it may in fact be quite wrong at places? For reference, Levavasseur and co-authors have provided some thoughtful contributions about downscaling in palaeoclimate applications (e.g.: Levavasseur et al., 2010, The Cryosphere, 10.5194/tcd-4-2233-2010, and subsequent references).

Yes, this is the approach we have taken, following Osborn et al 2016. We agree that there are limitations applying this approach to paleoclimate. These limitations were largely the reason for our choice of timeframe, as the assumptions would become increasingly untenable in deeper time, for instance as tectonic uplift progressively impacts on the validity. We have included a discussion of the weaknesses of the approach in the new section 9 (above)

Minor comments
L. 280-295: Define the R-squared score

Text added:
where R_c^2 is the coefficient of determination of the emulator of principal component c, evaluated under leave-one-out cross-validation of all simulations

clarify what is mean by "the performance averaged over the eight emulators" (since the P-metric is near one, an arithmetic mean of P may be inadequate).

The arithmetic mean is not required and text revised to:
The power exponential was found to give comparable or better performance compared to the other covariance functions in all eight emulators and was therefore chosen as the default covariance function, and used in all analysis that follows.

Clarify the approach implemented for calibrating the length-scales appearing in covariance functions: do they vary across components, across variables?

Text added:
We used an anisotropic covariance function (different length scales for each input dimension) and estimated the unknown length scale parameters using the type II maximum likelihood estimators (Rasmussen and Williams, 2006).

The original reference for Berger and Loutre, as quoted here, is Berger and Loutre (1991), doi 10.1016/0277-3791(91)90033-Q . It could be cited along with the Pangea reference.

Added

Figures 3 and 4 are a bit overwhelming, with small character size, and only the one with eyes trained in deciphering PMIP-type experiments will understand that the anomalies seen here are reasonably expected from GENIE-PLASIM and understand its limits.

We have removed the 2-component analysis to simplify the presentation, and have increased font sizes for improved clarity.

0.2 Conclusion
The article could be a nice addition to current efforts in simulating the Pleistocene climate, but it is ambiguous as to its objective. If the authors ambition is to provide an technical, significant improvement on emulation, then they need to focus more on the evaluation of the emulator as such, and be more thorough in the discussion of the different technical options. If the ambition is to provide a final product to be used by non-climate users, then I would urge the authors to be much more critical about the pitfalls of the current reconstruction, and in the present state, I would actually discourage dissemination of this product, since the risks of it being misused are too large.

We hope that the new Section 9 has addressed this important concern.