Reviewer 1

The STITCHES algorithm presents a unique time-sampling based approach that enables exploration of different, arbitrary climate scenarios. Its added benefit of not being limited to specific climate variables or spatial/temporal scales makes it a powerful tool in comparison to existing simple climate models/emulators. Overall, it is extremely relevant to the climate modelling and impact/integrated assessment societies and suitable for the Earth System Dynamics journal.

Thank you for your positive reception of this work and your careful and constructive review.

Some comments are as follows:

High-level comments:

1. The “outside the lower-end emission scenario bracket” application of STITCHES should be clarified, there is discussion surrounding overshoot however not for low-emission scenarios with near equilibrated climate by 2100.

We have modified our discussion throughout to include the lower end (or in general, extrapolation outside of the existing envelope) as a show stopper for STITCHES, together with the emulation of scenarios whose shape is not well represented at this time in the CMIP6 archive we are using as our sandbox. While emulating scenarios above the highest is definitely impossible for STITCHES, a scenario that is lower than the lowest available but still greater than or equal to historical levels of warming could be in theory emulated, but the meaning of such scenario could be argued as not exactly apparent.

We have added/reworded the last sentence of the abstract on these points: *Given that by definition STITCHES cannot emulate scenarios that result in GSAT trajectories outside of the envelope available in the archive, neither can it emulate trajectories with shapes different from existing ones (overshoots with negative derivative, for example) the size and characteristics of the available archives are the principal limitations of STITCHES deployment. Thus, we argue for the possibility of designing scenario experiments within, for example, the next phase of the Coupled Model Intercomparison Project according to new principles, relieved of the need to produce a number of similar trajectories that vary only in radiative forcing strength, but more strategically covering the space of temperature anomalies and rates of change.*

2. Some discussion on choice of tuning parameters (X and Z) for different temporal scales (annual vs monthly) should also be given. Since non-linear warming could

manifest more strongly at monthly timescales (due to e.g. snow-albedo feedbacks), this could limit the values of X or Z to be used (or otherwise the fineness of temporal resolution). Given that decadal oscillatory patterns such as El-Nino are aimed to be conserved, implications of having X>9 and the compromise this has on fidelity of representation for finer temporal resolutions should furthermore be explored (e.g. looking at performance on monthly timescales with different X values).

Perhaps we are misunderstanding the reviewer's point, here, but we think that monthly behavior would not be affected if not at the seams by a different choice of X and Z, given that once the sequence of pointers is created, the behavior of monthly variable is that of the original ESM output. The X and Z parameters apply to annual global temperature time series by construction, importantly because we want STITCHES to emulate scenarios on the basis of a trajectory of GSAT produced by simple models, which usually do not produce monthly output. Thus, X and Z, rather than reflecting on the behavior of monthly time series, are designed to ensure that what we are emulating is the forced component, and that we do not introduce severe discontinuities at the seams vis-a-vis the behavior of slower (multi-annual) modes of variability. Post-facto we do not encounter many cases where the behavior of monthly variables shows artifacts from the stitching, as documented in the validation section of our paper. We have added however a sentence to the section looking at the choice of Z that mentions the possibility of considering that for values of X very different from what we use, 9. WE invite the users to do that, as we provide the software where both values are tunable parameters.

3. Although discussion of application of STITCHES is given, readers would be curious for more discussion on future developments and improvements that could be made.

We think we can address the reviewer suggestion both by pointing at possible developments of our algorithm itself (for example, alternative choices of metrics for the nearest-neighbor space and distance) and importantly about what we see as promising developments in the field, with plans to join forces with other type of emulators, like MESMER-M and MESMER-X and a recently submitted emulator proposal, called PREMU.

In particular, within the last section of the paper we have added a sentence related to possible modification of the technical aspects:

Last, some technical aspects of our algorithm will benefit from further analysis/considerations: possibly some applications may be able to relax the tolerance parameter, and thus set the conditions for easier matching and more numerous stitched realizations. This might be true of applications that would not be too sensitive to interannual differences. On the contrary, tightening the tolerance to match specific ESMs' internal variability will be beneficial in eliminating spurious behavior that we have documented in some cases, especially when the archive of available runs is poor. *More generally we could choose a difference distance measure in the $(T,X*dT)$ space, or a completely different space over which looking for nearest neighbors, but the necessity of conforming to what a simple model can produce on the basis of a new emission scenario needs to be kept as a consideration.*

We have also concluded the paper with an explicit call for using the novel emulators that are being developed of late in a complementary manner:

The deployment of STITCHES, in concert with other emulators like MESMER-M and X~\cite{beuschetal2020,beuschetal2021,Quilcailleetal2022} and PREMU~\cite{Liuetal2022}, which are intended to produce new realizations of internal variability could then complement and enrich the effort of the ESM community.

With a new citation for a paper in discussion at the moment proposing a new emulator for precipitation:

Liu, G., Peng, S., Huntingford, C., and Xi, Y.: A new precipitation emulator (PREMU v1.0) for lower complexity models, Geosci. Model Dev. Discuss. [preprint], https://doi.org/10.5194/gmd-2022-144, in review, 2022.

Below are more specific comments

Specific comments:

L4: the link between emulators and computational demand should be clarified

We have added a few words here in the abstract to this effect: Given the computational cost of running coupled Earth System Models (ESMs), *which are usually the domain of super computers and require on the order of weeks to complete a century-long simulation,* only a handful of different scenarios are usually explored by ESMs. An effective emulator, *able to run on standard computers in times of the order of minutes, rather than days,* could therefore be used to derive climate information under scenarios that were not run by ESMs.

L19: This may be confusing to readers: the use of GSAT to create the pointers from which all other climate variables at different spatial and temporal scales will be stitched together should be clarified (i.e. pointer is not climate variable specific).

Thank you for underlining this, it is really the crux, and we will make sure to clarify, also given the comments of Reviewer 3 which indicate the need of being more careful with the use of our terms and language, evidently confusing to some. In the abstract we have reworded now by specifying: A look-up table is therefore created of a sequence of existing windows/experiments that, when stitched together, create a GSAT trajectory "similar" to the target. *Importantly, we can then stitch together much more than GSAT from these windows, i.e., any output that the ESM has saved for these experiments/time windows, at any frequency and spatial scale available in its archive.*

L113: This suggestion is a bit strong given that emulators already mentioned (Link et al. 2019, Beusch et al. 2020,2021) circumvent the need for initial condition ensembles by providing stochastically generated imitations of the expected internal variability. Furthermore, scenario exploration to look at climate under equilibrated or overshoot state is still extremely important, and this should be clarified.

Absolutely agreed, and we have reworded this sentence altogether as: *If emulators, possibly used in a complementary fashion, become part of the overall strategy in providing climate information to the impact research community we argue that the next ScenarioMIP design may identify different priorities from the current one.*

L115-L135: Very well explained background to the rationale!

Thank you.

L146: what about scenarios lower than the lowest emission scenarios or overshoot scenarios?

We rephrased adding : We note here, however, that by construction our algorithm does not allow extrapolating to levels of warming above those of the highest scenario available in the archive, *or below the lowest.*

We initially did not worry about lower than the lowest, since the historical simulations would be the lower limit, and those are lower than the lowest, and available. Realistically though interesting scenarios lower than the lowest would be overshoots, and for those our caveats about the lack of a rich-enough archive remains valid. We discuss this latter point later on, after introducing the (T, X*dT) space. We write: Note that when the goal is emulating non-existing scenarios, our targets need to be trajectories that reach warming levels within the ones available as building blocks in the archive, as our algorithm does not allow extrapolating. Similarly, STITCHES stops short of being able to emulate overshoot scenarios, given that the archive does not offer a large population of overshoot experiments that we can use as building blocks (i.e., the cooling behavior of GSAT in an overshoot experiment cannot be sampled from increasing, or flat, GSAT trajectories). These considerations could be useful to keep in mind when designing the next phase of ScenarioMIP.

L197-L205: Z is dependent on X which is also a tuning parameter, this may introduce additional caveats in choosing X so as to avoid "jumps" between the seams. Have sensitivity tests been performed on this? Some explanation on how to jointly pick the optimal combination of X and Z should be provided.

We have kept the two choices separate, as we worry about X in the context of adapting the smoothness of the archived/available GSAT series to the time series that wewould get from a simple model. We then have a session later in the paper that discusses our investigation of the sensitivity of the algorithm results to the choice of Z. Please see Section 3.1.3. Our goal is to publish the code where all these parameters can be tuned (to specific ESMs, and specific applications) rather than trying to come up with gold standards that we believe would be anyway sensitive to the two choices mentioned above. We do add a sentence however, in this section, inviting exploration of the choice of Z, depending on values of X.

L211: Is the ensemble size the sole thing considered when choosing which ESMs to display? Looking at ESMs of different genealogies would also be interesting especially for the (T, XdT) space (if not that is also O.K., just curious about why the above criteria).

We chose to develop our emulator on the basis of the CMIP6/ScenarioMIP archive and are using all models that provided a subset of monthly and daily variables that we set out to emulate. Some of these models have very small ensemble sizes, some have large ensembles. It would be possible to look at past generations of models. We wonder if the reviewer is thinking about combining the archives across the same model's different versions. That, we think, would be problematic given how different successive versions of the same model can be. So we did not go there. Ideally, the same version of the model would have run both sets of scenarios and that would make the archive richer, but we have not found that to be the case, with most ESMs having submitted a new version to the latest phase of CMIP.

Figure 1: it seems that for most models around -0.01degC the rate of historical warming is higher than that at 0-0.01 degC, is there a reason for this? It also raises the question of the generalizability of this approach for time windows with major volcanic events (e,g, Mt Pinatubo which has a distinct fingerprint in the GMT trajectory) and some elaboration on this may be required.

The reviewer has identified something that we did not notice, having focused our application to the scenario part (future) of the simulations and that indeed seems specific to volcanic eruptions. We have added a sentence in the conclusions pointing this out:

There are more subtle aspects of stitched scenarios that may pose questions of fidelity and representativeness. *We have not addressed the challenges that short but intense forcing episodes, like volcanic eruptions, may pose, since we have focused the application of STITCHES on future scenarios, which do not represent them. A careful look at Figure~\ref{fig:PANGEO_archive} can highlight a region of the space populated by grey dots (the historical part of the simulations) showing a peculiar pairing of absolute temperature anomalies and rate of change in the region around $T=-0.01$ compared to that around $T=0.01$. This would suggest a specific behavior of GSAT while recovering from volcanic eruptions that is not easily emulated by finding analogs in the historical period (away from volcanic episodes).*

L227-L230: Great that this is elaborated upon here! Providing this elaboration earlier could benefit and provide more structure to the text however.

We have added these points to the Introduction. Specifically, we added a sentence in the next to last paragraph:

We split the GSAT trajectory into regular windows, and we identify for each of them a "nearest neighbor" among the windows of GSAT trajectories available in the archive, from the finite number of experiments that were run and archived, *as long as the scenario that is target of our emulation is characterized by an intermediate level of GSAT warming, and similar rates of change to those present in the archive.*

And we pointed out explicitly overshoots and stabilized scenarios in the last paragraph:

We also discuss the challenges that STITCHES encounters when targeting scenarios of shapes other than regularly increasing forcings, *like stabilized scenarios and overshoots,* therefore suggesting that a concerted effort in exploring scenarios of different shapes, rather than scenarios that only vary in the strength of the radiative forcing, could be made to facilitate the application of the emulator.

Figure 2: It seems that all ESMs in this figure have a mismatch in the GSAT trajectories after 2050 for ssp 2-4.5 (and also BSS-CSM2-MR and CMCC-ESM2 in Figure 4), some elaboration on this may be needed e.g. transient vs equilibrated state. In general some consideration of how to stitch together cases where X*dT ~ 0 should be elaborated as nearest neighbors could have both a positive or a negative trend.

We have added a sentence, as suggested:

Also from these figures one can assess that the behaviors that appear to deviate from the expected, are all at the tail end of the simulations, and only for those models that offer only one pair of scenarios in the archive to sample from, *particularly for SSP2-4.5, which adds the extra challenge of a trajectory that stabilizes ($ dT \approx 0$) and needs to find matches among windows from scenarios that, at that level of warming, are by construction increasing in forcings. In general, stabilization scenarios together with overshoots pose a challenge to STITCHES given the content of the CMIP6 archive from which we construct our emulations.*

L306: It would be interesting to see month specific trends (e.g. the decadal trend for Jan and Jul). It seems here it is only the decadal trend of the whole monthly time series, if not this should be clarified as well.

We will provide these.

Figure 6: There seems to be systematic overestimation of monthly variance around central Africa (also for models in the appendix), are there reasons for this (e.g.

vegetation/land cover changes where SSP 5-8.5 imposes quite high deforestation which may lead to spurious variabilities)

First, we realized we had by mistake included a panel for this monthly TAS variability plot in the appendix for CAMS, rather than MIROC6. MIROC6 SSP2-4.5 does not show the same patches of overestimated variability over central Africa, while only the lower area is present for SSP3-7.0. It may be true that some effects of vegetation may be surfacing here but it would be fairly speculative of us to discuss this, also given the fact that, for CAMS, the pattern is the same for 4.5 and 7.0, that have different land use assumptions.

L321: The argument that internal variability explains the mismatch in the Arctic is not so convincing. It could for instance be due to the AMOC or otherwise due to a non-linear increase in summer time temperatures during ice-free arctic summers.

We identified internal variability as the explanation because the patch appears in two out of three ensemble members, but we are happy to add these other explanations as possibilities, as suggested:

Internal variability is likely responsible for an area in the Arctic appearing as inconsistent in two of the three realizations, *but effects of ice-free summer intensified warming, or behavior of the AMOC could contribute to this limited area of disagreement.*

L346: Figure 7, it may be difficult to visually gauge similarity in magnitude and oscillatory behaviour. Although this is made more obvious in Figure 8, it may be a good idea to apply a power spectral decomposition instead and show their results for a clearer overview. Very good idea to look at SOI within the analysis otherwise!

We will include spectra.

L400: Does the Z_cutoff value generalize to all values of X? The calculation of Z_cutoff is already a very useful exercise so this is a minor detail, just curious.

We haven't gone there but added a sentence pointing at possible exploration of this issue, enabled by our software, in Section 3.1.3 (about the sensitivity to choices of Z of the number of ensemble members obtainable).

L438: The term envelope collapse should be clarified and how it related to the Z value as well (i.e. how best to know at which Z envelope collapse has been approached?

We have rephrased the sentence where the term appears in the Methods section (step 6) to clarify its meaning and also added clarification to Section 3.1.3.

Table 5: Is there a relationship (e.g. linear) between between E_r and Z_cutoff, or are they stable and then jump to above 10% after a certain cutoff?

For a particular ESM, E_r and Z_cutoff tend to increase together. However, due to the discrete nature of our matching set up between a finite number of target and archive windows, there are clear stable values followed by step increases in this relationship. Many values of Z can result in the same set of archive windows matching to a target window, until Z crosses some threshold and another archive window gets added to the set of matches. At a certain point, Z increases enough that the next added archive window is too different from the target window to be a 'good' match (and E_r has a step increase to reflect that). This, combined with the fact that the specific generated ensemble members for a given Z value are stochastic, is why we select Z_cutoff via this post-hoc set of experiments rather than directly within the algorithm itself.

Table E1: The E_1 and E_2 values for CanESM5 tend to be higher for 20 archive members and then drop lower at 25 archive members. More so for SSP 3-7.0 the E_1 values are 0 at 25 archive members for both 2010 and 2050. Is there a reason for this?

We regard this table as showing fairly noisy results…we are submitting STITCHES to a tall order in having to emulate two scenarios multiple times on the basis of two bracketing, very different scenarios. The way we set up the exercise is by randomizing the members of the full archive (of CanESM5 in this case) included in the smaller ensembles (e.g., we choose 5, 10, 15, 20 members randomly from the 25 available). The algorithm also randomizes the choice of nearest neighbors. So, the patterns of these metrics are not easily interpretable. We would expect that if we repeated this exercise many times the average outcome would lend itself to a better interpretation, but this exercise is mostly about showing the strain imposed on the algorithm when supplying such extreme brackets.

We have added a sentence that highlights the noisy nature of these results when pointing at the table:

We have performed the same exercise by limiting the archive to the two bracketing scenarios, SSP1-2.6 and SSP5-8.5, and trying to construct ensembles for SSP2-4.5 and SSP3-7.0. In this case STITCHES is significantly challenged, and its performance as measured by the E_r metric significantly diminished *and, when comparing what happens for the same model and increasing numbers of archive members, unpredictable, due to the fact that the algorithm randomizes both the identity of the archive members and the choice of the nearest neighbors to construct the emulated output.*

Conclusion and Discussion: the recommendation for looking at less scenarios and focusing on more initial condition ensembles may be quite strong: perhaps there should be elaboration on which scenarios are more useful to explore (i.e. ones where interpolation becomes difficult such as overshoot or equilibrated climate). The applicability of STITCHES across different temporal scales should also be clarified (i.e. limitations when applying it to annual vs monthly vs subdaily timescales).

We have modified our discussion of the implications for CMIP/ScenarioMIP by simply pointing at the need of populating the space of (T, dT) more effectively. We are also calling for the use of other type of emulators jointly with STITCHES.:

The next phases of CMIP could complement what is available now by deliberately exploring types of scenarios that are not well represented in the current archives, like stabilized trajectories and overshoots. The challenge would lie in choosing the best set of runs to optimally populate the $(T,X*dT)$ space to maximize the number and shape of attainable new trajectories from the existing ones. The deployment of STITCHES, in concert with other emulators like MESMER-M and X~\cite{beuschetal2020,beuschetal2021,Quilcailleetal2022} and PREMU~\cite{Liuetal2022}, which are intended to produce new realizations of internal variability could then complement and enrich the effort of the ESM community.

Editorial comments:

L35: support the climate information needs of the impact research community

L44: bias-correcting them. Alternatively just bias-correction could also work

L120: perhaps “scenario-independence” would be a term more consistent with the terms already introduced

L147: “the STITCHES algorithm”

Figure 1: Lovely plots, very informative! Font size needs to be increased however.

Thank you, we have adopted these edits and the new figure will have larger fonts.