Reviewer 1

The authors have revised their manuscript and thoughtfully considered my comments and those of the other reviewers. In particular, the authors have made efforts to clarify their framework and added statements on its benefits and weaknesses. This is an important study which will hopefully initiate more discussion on climate modelling efforts in our community- congratulations!

We thank Reviewer 1 for the thoughtful original comments and the positive reception towards our revision.

Reviewer 2

I believe that the author's have answered most of my questions sufficiently and improved on the technical quality as well as readability of this paper. There are some areas that could still be elaborated/improved upon as follows:

L12: the term "construct ESM output" is misleading as you are constructing "ESM-like output" or otherwise "imitating/emulating" ESM output
Fair point, we reworded as ESM-like output, as suggested.

L101: when referring to Beusch et al. representing higher frequency fields, a more suitable reference is: https://doi.org/10.5194/esd-13-851-2022, same goes for L207 and L650
We have substituted (first and last instance) and added (middle instance) this reference, thank you.

Generally piece, segment as well as piecing and stitching seem to be used interchangeably and this could be streamlined.
We have now tried to be consistent throughout. In particular, we have substituted 'pieces' for 'segments' in all cases (4) but for one instance where we say:
*Our algorithm is applied separately to each individual ESM, as stitching together different models' **lengths of simulations** would almost certainly introduce spurious behavior.*
We have only found one instance of the use of piecing, and we realized we were using it in an incorrect sense. We have now changed that into "splitting" as in:
*the same smoothing and **splitting** procedure is applied to the trajectory of GSAT for the target scenario*

L210-L255: from reading other reviewer's comments I can imagine this is hard to follow for the general ESD audience. Figure A1 enriches this but could be made clearer. Some suggestions would be to put a legend for what the blue square vs yellow circle represent as well as numbering steps within the figure according to their corresponding step as outlined in the text.

In general, the steps in Figure A1 after the first middle blue box is hard to follow or match to the text and could be made clearer.
Thank you for the feedback. We have followed both suggestions: we have now a legend at the bottom of the diagram explaining the difference between yellow circles and blue boxes and we have added numbering to the side of the boxes that refer to the respective steps as detailed in the itemized list of the "Methods" section.

L340-L350: I appreciate the discussion of higher frequencies stitched together having a lower likelihood of introducing GSAT trajectory "jumps" due to high noise within the introduction section. The fact that this is being demonstrated here should be made more explicit however, additionally what are the GSAT's recovered from? monthly or yearly gridded temperature?
We have added a parenthetical clause to clarify that our GSAT time series are made of annual values. The paragraph now starts as:
For all cases when the emulation of GSAT time series (made of annual average values) […]

We have also added a short paragraph at the end of the section following the reviewer suggestion to stress the results of this validation of higher frequency, noisier quantities:

On the basis of these results, we confirm the correctness of our expectation that, after validating the statistical characteristics of a large scale, low frequency quantity like annual GSAT, further validation of emulated variables at grid-point scale and higher temporal frequency do not seem to present larger challenges. The higher noise of these quantities indeed accommodates the discontinuities introduced by their emulation.

L415: a good reference for ice-free summers could be: https://doi.org/10.1175/JCLI-D-15-0284.1
We have added this reference, thank you.

Figure 5: The term monthly trends remains a bit ambiguous here: is it a month-specific trend, in which case the month should be specified in the title/caption. If not how was this calculated? Since e.g. winter monthly warm quicker than summer, one would expect that this would be taken into account here and if not why?
We have added the following text when describing these results:
All metrics here are computed using time series of gridded output at monthly frequency, covering the entire annual cycle, for the length of the emulated output (2015-2100). In the appendix we show similar results for month-specific output sampling behavior during Boreal winter (January) and Boreal summer (July), addressing the possibility that the emulation could be differently challenged by stronger or weaker forced trends.

Figure D7: nice addition! *Thank you*

We thank reviewer 2 for the thoughtful original comments and the positive reception of our revised version, together with these additional points for improvement that we hope to have addressed as detailed above.

Reviewer 3

I thank the authors for the clarifications they brought after my admittedly candid comments on the original manuscript. I now have a better feel of the STITCHES procedure.

We thank the reviewer for giving the study another try. We feel as if there is still a fundamental challenge in the communication of our algorithm's purpose and uses. We have tried to respond to the reviewer's comments here, but we could not really find a justification to modify our text and figures, as a consequence. We also think some of our text and figures have been misinterpreted but given that the other reviewers do not raise issues, in fact in some instances are openly happy with our figures we would ask not to be made to change what we have included.

Major comments

I am a bit skeptical about the interest for climate scientists of the paper. Unless I missed it in the discussion section, there is no mention of internal climate variability that makes ensemble members (of the same model) yield very different behavior (e.g. Deser et al. Nat. Clim. Chang. 10, 277–286 (2020). https://doi.org/10.1038/s41558-020-0731-2). I think that the physical caveats should be discussed in the light of the literature on climate variability, otherwise I feel that results reported in Figure 2—4 are likely to bring overconfidence in climate projections to impact communities.

We think that the interest by climate scientist in our solution will come from its potential in lightening the burden of running many similar scenarios, a demand that comes from the impact modeling research community. Please, see later for more on this point.

As for STITCHES and its emulation of internal variability (IV), that is indeed a very important aspect of emulation that we seek to address, when we set out to emulate high frequency ESM output. Were we not concerned with IV, simple pattern scaling of multidecadal mean behavior could still be used as a solution to emulation. However, as we describe in the paper, state of the art impact models require realistic sequences of relatively high frequency quantities, and modelers are aware of the need of addressing the interplay of forced changes and IV. Hence the need for an emulator that preserved aspects of IV, besides the forced component of climate change.

In the paper, after validating aspects of IV specific to a single realization we specifically validate the emulation of initial condition ensembles. This is one of two specific uses of the emulator, and we have dedicated a subsection to its validation. We introduced the metric Er, one component of which, E2, addresses the accuracy with which the behavior of the emulated ensemble mimics that of the real one. Table 4 and Section 3.1.2 describe these results.

Also, by construction, the algorithm is careful not to repeat pieces across the initial condition ensemble members emulated, in order not to create identical behavior across ensemble members.
Last, our algorithm replicates, again by construction, the internal variability of an ESM within the pieces used in the stitching. Of course, also by construction, and as we discuss, characteristics of internal variability that go beyond the 9-year window used for our pieces (i.e., constituting the building blocks) won't be emulated faithfully, and we clearly describe this among the caveats, when adopting STITCHES output rather than true ESM output.

One un-discussed issue of the paper is climate sensitivity: since STITCHES is based on GSAT variations, one could in principle test time series of CO2 atmospheric concentrations, which are available in ESM simulations with ssp scenarios. Is it possible that STITCHES violates the underlying climate sensitivities for each model? This should be fairly easy to assess, and might interest the communities who deal with biogeochemical cycles.
We apologize, but we do not understand what the reviewer is proposing here. We show in our validation section that trends, specifically trends in temperature, and century-long trends in particular, are preserved for the individual models emulated. That would suggest that STITCHES does not modify/violate the climate sensitivity of the emulated ESM in any significant way (again, for the purpose of impact modeling). We are not sure how impact modeling would be concerned with climate sensitivity per-se, beyond the need to choose ESMs that span its assessed range, which the CMIP6 models used here, and against which we validate our algorithm successfully, surely do.

Specific comments
I am still puzzled by some statements and figures of the revised manuscript.
Figures 2-4 do not look very informative, at least from the statistical point of view, as the lines all overlap each other, and all look undistinguishable. Is there a case where stitched and target time series do not look alike? Could all the panels be summarized in just one figure? (See my first major point)
We do not intend these figures to substitute for a rigorous validation, which we describe extensively in the text. We wanted to present the reader with the range of models/ensembles that we set out to emulate, and give a visual impression of the results, leaving to the text the details of our extensive validatory procedures. As for cases where stitched and target time series do not look alike, it is generally a result of selecting a Z value (tolerance for the search of nearest neighbors) that is too large, and represents an emulation failure. Rather than plot the results of failure, which may take many forms, we devote a section of our results to documenting our estimation of Z_cutoff, the value at which the generated ensemble is statistically consistent with the target ensemble and therefore appear 'indistinguishable' (i.e. clearly following the same trend as the target and displaying consistent internal variability).

The ENSO section is strange and very qualitative: from the described procedure, would there be a possibility that the emulated versions DO NOT look like the original versions? In other words, is it not by construction, from the reshuffling procedure, that ENSO time series yield

realistic features, at least visually? The auto-correlation figures (Fig. 8) are not particularly convincing. In particular, those graphs do not show frequencies, as claimed in the text (l. 397). Why is the auto-correlation (ACF) equal to 0 at lag 0 for the CAMS model? The ACF at lag 0 is the variance, and should NEVER be zero!

*We wanted to once again give a visual impression with the first type of plots, and then substantiate our claims with the temporal and spectral characteristics of the time series (we added a figure for that purpose following Reviewer 2's suggestion, which they appreciated). The 0 values at lag 0 are seen in the plots of the PARTIAL autocorrelation functions (**P**ACFs, second and fourth rows of plots) not of the autocorrelation functions (ACFs, first and third rows of plots). The PACF is defined as the partial autocorrelation of the values of a time series with its own lagged values, and therefore is defined for lags 1 and higher. The plots of the ACF show the taller spike at lag 0, as is to be expected, indeed representing the variance of the time series. Also, we are not claiming that the ACF/PACF graphs show frequencies. Assuming the reviewer does not agree with the use of the term "densities" we are referring to the additional figure D7 in the appendix, which shows spectral densities (requested and now approved by Reviewer 2).*

Could the authors give an illustrated example of the utility of STITCHES to climate scientists? (see my second major point).

STITCHES is a tool conceived to support impact modeling, first and foremost. Its utility to climate scientists resides in "freeing up" the resources that are right now dedicated to running many slightly different scenarios, or many members of initial condition ensembles. Particularly, our figure 1 highlights a view of ESM outputs that may be valuable for CMIP7 scenario planning as climate scientists consider trade-offs among allocation of computational resources to different type of experiments (scenarios and other types of simulations). We do not claim that our emulator has value in uncovering physical behaviors that a climate scientist would find novel. What we confirm from our validation exercises has been well known for the last few years at least: many impact-relevant variables are "slaved to GSAT", and not path dependent. This is an interesting result from the climate science research community that we exploit in patching together our new scenarios/ensemble members on the basis of the corresponding behavior of GSAT. Thus, we could claim that STITCHES allows the climate science research community to benefit from their own finding in a concrete, resource-preserving way in future exercises.