Response to Referee #2

I read the manuscript only regarding the validity of the used statistical methods, but not regarding its application domain content. I found the overall method and details convincingly motivated and could not detect any mistakes.

Thank you for this comment and for reviewing the method.

I have only one request for improvement:

Line 212, "which is close to...": In view of the corresponding remarks by anoner reviewer, please replace this statement by a short discussion of the results of a binomial test where you *test* the hypothesis that the underlying probability of containment equals the expected 90%. I expect that the relative frequency of only 85% observed in your leave-on-out exercise will constitute a *significant* deviation from the expected 90%, at least if you assume all 200 (is this number correct?) trials to be independent. If you choose a smaller number of degrees of freedom because you consider the trials to be partially dependent, please justify your choice of degrees of freedom in that binomial test.

As suggested, I have conducted a binomial test with a chosen degree of freedom of 27. Although the number of pseudo-observations is 172 (see table A1), we cannot consider the members of the same model as completely independent from each other as they share the same forced response. Therefore, I chose to apply the assumption of independence to the models rather than the members. This choice is also debatable since many models share common components; nevertheless, it is much less harmful than considering the members of the same model as independent. In practice, for each model, the number of times the pseudo-observation is contained in the constrained range is weighted by the number of members of that model (e.g., by 1/50 for CanESM5, see caption of Table 1). In this framework, the binomial test indicates that a success rate of 80% remains compatible with an expected rate of 90% (p-value = 0.18). I have added this point to the manuscript.

Minor questions and corrections:

L.21: "has" --> "have"
L.41: "in this paper" appears twice

Fixed

L.45: Please explain shortly the motivation for calling this "Kriging" and the relationship to ordinary or Bayesian Kriging since that is nonobvious.

This term is used to refer to the interpolation aspect between observations and models. I added this clarification to the manuscript.

L.54, "median": I welcome taking a robust statistics approach here. But when taking the median rather than the mean, should one then not also take the root-mean-square (or even

mean absolute) difference from the median (!) as the corresponding estimate of variability? and how could one then treat covariance in such a robust statistics approach?

I do not fully understand this question. $\Sigma_y = \Sigma_{meas} + \Sigma_{iv}$ is the observation error covariance matrix, where Σ_{meas} and Σ_{iv} describe the measurement error and internal variability, respectively.
Σ_{meas} is estimated as the sample covariance matrix over the 200-member ensemble of the HadSST4 dataset.
Σ_{iv} is estimated using observed annual SST time series over the 1850-2021 period. I do not use the observed median to estimate the forced component, but the CMIP6 multi-model mean which provides a priori a better estimate because as you seem to say (at least as I understood it), the trend in the observations is polluted by internal variability. Hence it is difficult to separate the two components. Both errors due to the internal variability and the measurement uncertainty are taken into account in the calculation of Σ_y.

L.56, "the HadCRUT5 ensemble of 200 members": Are these 200 members in a one-to-one matching correspondence to the 200 HadSST4 members, so that covariance between SST and GSAT can be estimated?

The HadCRUT5 and HadSST4 ensembles have been generated independently, hence there is no relationship between the members.

L.98: "a some" --> "some"
L.112: "are to" --> "are used to"

Fixed.

L.150: "half of the observed cooling": Why is this so low, is the internal variability that large? Maybe point out that at least the confidence intervals of "Obs" and "constrained ALL" overlap.

L.183: "warmig" --> "warming"
L.193: "confidence these" --> "confidence in these"

Fixed.

L.200: "Sigma_y..." --> Is this done in the same iterative way (D3) as for the actual results?

Yes, I have added this precision.