We are grateful for this thorough, constructive and very useful review. Our responses are in italics

Interactive comment on "PLASIM-ENTSem: a spatio-temporal emulator of future climate change for impacts assessment" by P. B. Holden et al.
D Wilkinson (Referee)

I enjoyed reading this paper by Holden et al. It demonstrates the type of statistical analysis that can be used and is necessary to understand the range of response that can be obtained from climate simulators. While computational resource is limited in comparison to the complexity and computational demands of the models, we will need to use statistical tools such as emulators to do the types of inference we are interested in.
The paper uses advanced statistical methods in innovative ways in order to cope with the complexity of the simulators. They combine emulation with a form of principal component analysis, to show how we can predict future climate in response to arbitrary radiative forcing scenarios. The methods show how we can overcome some of the computational challenges and constraints that are constantly faced in climate science. I like their use of diagnostics, and the Chebyshev polynomials to reduce the radiative forcing to a lower dimension, and their use of principal component emulators.
However, I have some concerns about some aspects of the analysis which I detail below. While I think these points are important, and would improve the analysis, I think the method as it is still demonstrates the potential of these statistical tools.

Concerns/suggestions:

1. Use of non-centered data in the PCA. In section 4.1, a dimension reduction method is introduced which is based upon EOFs/principal component analysis (PCA). When doing PCA, it is usual to decompose the covariance matrix using the singular value decomposition of the centered data matrix. Here, the authors prefer to use the non-centered data. This is no longer PCA. It may still be a useful decomposition, but I think there are dangers to using it, particularly in its interpretation.
If we use (centered) PCA, then as the authors say, the first component will be the direction in the data about which there is maximum variation. If we use the authors' non-centered decomposition, this will no longer be the case. As illustration, imagine a cloud of points not near the origin which are roughly scattered along a line. Centering the data will move the cloud so that it surrounds the origin, and then PCA will find the direction of maximum variation, i.e, the line the points lie along. The next component is constrained to be orthogonal to this, and in the direction of the next greatest variation etc. If we use uncentered PCA, then the first component will be a line to where the cloud is in space. Later components then have to be orthogonal to the first component, and may thus no longer have any sensible interpretation.
Subtracting the mean in PCA removes the first order location information from the data, and leaves the principal components to explain the second order variance/covariance information. If we use non-centered PCA, then the

components have to capture both first and second order information. The authors rely on the percentage of variance explained to justify using a non-centered decomposition, but this could be misleading. The first EOF in non-centered PCA just moves us to the data and is a measure of location not scale, so although its eigenvalue may be 93% of the sum of the eigenvalues, this does not mean it explains 93% of the variance in the data.
The non-centered PCA does provide a decomposition of the data, and it may well be that it is a useful approach. However, the authors need to demonstrate this by showing that their emulators reproduce the spatial fields well after reconstruction, not just that they can reproduce the PC scores (which is what is done at present). I also think that "PCA" is a bad choice of terminology for the uncentered method, as it is no longer principal component analysis. They should also remove some of the description of the method as PCA (e.g. lines 9-10 on p 3361).

We agree with the referee and will address these concerns in the revised article, reconsidering our terminology (and approach if necessary) and quantitatively examining the extent to which the emulator is different when built from centred and uncentred data. As suggested, we will quantify the performance of the emulators in terms of their skill in reproducing the individual simulated fields, rather than just the PC scores.

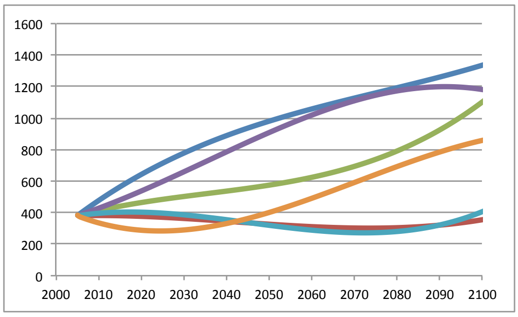
2. The authors choose to use linear models as parametric emulators, rather than non-parametric models such as Gaussian processes (GPs) as is more commonly done. There are some dangers to doing this.
(a) We are insantly limited to our imagination when building models, and the authors here only consider models which are quadratic or simpler in the covariates. Using a non-parametric model removes this limitation.
(b) In linear models, uncertainty is typically either ignored, or is represented as a constant band of uncertainty surrounding the fitted parametric model. This means that when we make predictions with a parametric emulator, the uncertainty in the prediction is always the same, regardless of whether the region of space we are making predictions in is one in which we have lots of model evaluations (and hence can be quite confident), or more importantly, in an extreme region of parameter space in which we have very little information and thus are very uncertain. The model has no mechanism to tell the user whether it fits well or not in this region of space, or whether its predictions are confident or not.
Although GPs can be difficult to work with in large datasets, that is not the case here, as generally each ensemble has less than 1000 members, which is well within the scope of GP models.

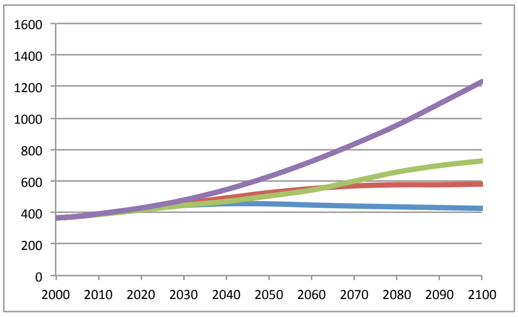
We agree that there are very good reasons for a Gaussian process emulator, as discussed in detail by the referee. We will expand the discussion on the relative merits of Gaussian process and parametric approaches. We note that our primary requirement here is to derive an ensemble of emulations that span the behavior of the simulator rather than to understand the uncertainty associated with individual emulations.

We note that the emulator is not used in extreme regions of parameter space. For each of the 188 emulations, we restrict ourselves to the parameterisations (in 22D parameter space) that were explored in the 564-member training ensemble. The only sense in which we use the emulator to explore regions of input space about which we have no information is in the use of different future forcing assumptions (in 6D chebyshev coefficient space). The 564 simulations used to explore this parameter space are forced with coefficients that are allowed to vary more widely than is likely in the likely real world future. To illustrate, the chebyshev ranges in the design ensemble are compared to the ranges of the RCP fits in the table below. This demonstrates that reasonable emissions scenarios are comfortably spanned in the design ensemble, so that emulations at the extremes of the 564-member sampled 6D space are unlikely.

	MLH Design (width)	*RCP fits (width)*
TC1	*0 to 1000 (1000)*	*51 to 930 (879)*
TC2	*-200 to 200 (400)*	*-58 to 162 (220)*
TC3	*-100 to 100 (200)*	*-15 to 22 (37)*

The two plots below demonstrate this further. The left hand plot illustrates the CO2 forcing profiles for the 6 training ensemble members that have either maximum or minima of the 3 CO2 chebyshev coefficients. These can be compared against the RCP CO2 profiles in the right hand plot, which are in general far less complex than profiles allowed in the training data.





3. The methods in the paper ignore the error in the emulators. Emulators are used as approximations to the simulators, and are used to rule out parameter values as implausible in the MPEF ensemble. If the uncertainty in the emulators is ignored, then we may end up ruling out regions of parameter space due to poor accuracy of the emulator. If we quantify the uncertainty in the emulator, then we only rule out parameter values as implausible if the emulator prediction interval (99% say) doesn't contain the observations.
Uncertainty could be incorporated crudely using the linear model approach used in the paper (while acknowledging the crude uncertainty bands that come with parametric models), or better still, Gaussian process emulators could be used to fully account for emulator uncertainty.

We agree and will discuss emulator error more carefully in the revision. Although beyond the scope of a revision (~10 CPU years were needed to perform the MPSF ensemble), this suggestion for an alternative design approach should be discussed. We note that our approach (using emulators in the design process in order to span *plausible parameter space as widely as possible) represents a substantial improvement on standard approaches such as a latin hypercube design, which we have shown, does not in general produce plausible climate simulations (unless the parameter ranges are confined to sufficiently narrow ranges about some tuned configuration).*

4. I wondered if the authors have considered separating time and space in the dimension reduction. At present, the authors collapse the time-series of spatial fields for each simulation, {y1, . . . y10} say, into a single vector yi, and then combine these into matrix Y = [y1, . . . y564] (of dimension 20480 by 564), and then decompose Y using PCA. They then build an emulator which maps from the parameter θ to each of the d leading principal component scores r1, . . . , rd: where β is the parameter vector.
$ri = \beta igi(\theta)+\varepsilon$
Instead, we could decompose the spatial fields only, collating the spatial fields from different times to give more observations. So that would be applying PCA to a 2048 by 5640 matrix Z = [y1, y2, . . . , y10]. Then the emulator would include time as a covariate and be of the form:
$ri = \beta igi(\theta,t)+\varepsilon.$
This is likely to be successful if the modes of variation in the spatial fields are similar through time (which seems plausible), and if the trend in time is able to be modelled as a linear model.
An alternative, would again to be to use a Gaussian process emulator with a separable covariance function in space and time. It is not clear that either of these alternatives would be better than what has been done, but they might be interesting to consider.

We did consider this approach, but preferred our approach for the same reasons proposed by the reviewer. By collapsing the time series, we allow for the possibility of abrupt state transitions (i.e. spatial fields are not forced to be similar through time). Additionally, we do not need to assume that a linear model in time is required. Consider the emulator as a function of the 3rd chebyshev coefficient. The relationship is certainly non-linear in time, being a combination of cubic and linear terms (Eq. 3).

We agree that this alternative is very worthy of consideration however, and will discuss it in the revised manuscript.

5. The paper would benefit from a clear mathematical description of what distributions are being approximated at each stage. For example,
• In Section 3, the Modern Plausible Emulator Filtered ensemble is (I think), a Monte Carlo approximation to
π(θ|D1, S1) where θ is the parameter vector, and D1 and S1 are the data and simulator
ensemble used. • The modern-plausible-simulator-filtered parameter set would then presumably be π(θ|D1, D2, S1, S2).
• The final paragraph in section 3 describes an ABC (Approximate Bayesian computation) algorithm, where parameters are drawn from a uniform prior

distribution, and then accepted or rejected according to whether the predicted simulator output is within certain tolerance limits of the observations. The paper might also benefit from a pictorial representation of all the aspects that have gone into the statistical analysis. This would make the analysis more transparent, and allow the reader to see clearly which simulators, datasets, and statistical assumptions have been used and where. At present, this is difficult to piece together from the manuscript.

We agree and will address these suggestions in the revision.

6. As a statistician, rather than a climate scientist, I found some of the most interesting details had not been included, presumably for reasons of space. For example, in section 3, the authors say a series of exploratory ensembles were generated, and used in some sort of screening process to determine what the important variables were. How to do this for expensive simulators is a non-trivial question, and it would be interesting to read the approach taken.
Also in section 3, the authors say the parameters were varied over their plausible ranges. It would be interesting to know how these plausible ranges were determined. If the ranges are too narrow, we risk missing important regions of space. Conversely, if the range is too wide we may miss interesting non-linear trends in the simulator output. The scale of the challenge faced is illustrated by noting that this 22 dimensional parameter space has 222 ≈ 4 × 106 corners, which is being explored with only 500 simulator evaluations, and so having sensible tight prior ranges would be a great help.

We agree and will address these suggestions in the revision. Our approaches were pragmatic. In essence, we considered the widest possible ranges over which we could physically justify varying the parameters and then ruled out some regions of parameter space in which it was clear that the marginal posterior for the parameter in question could never produce a plausible climate.

7. I could not understand the discussion at the bottom of page 3363 and the top of page 3364 (mentioned again in the conclusions) about the similarity of the distribution of the emulated and ensemble PC scores. It is unclear why noting this similarity is important. I thought that we wanted individual cases to match, which would automatically lead the distributions to match. Knowing that the distributions match does not automatically imply that the emulator is successful.

We agree with this and will discuss this in the revision. As noted in response to an earlier comment, we will revise the analysis to quantify the performance of the emulators in terms of their skill in reproducing individual simulated fields, rather than just the R^2 of the PC scores.

Our argument for this approach has been that we do not require individual emulations to reproduce the respective simulation. Rather, we need to produce an ensemble of plausible emulations for impacts. In applications to date, we have only used the emulated ensemble to derive a spatial description of the expectation of the climate change variable and its associated uncertainty. We will review this in light of the improved cross-validation.

8. The authors might consider citing Williamson, Goldstein, and Blaker 2012, “Fast linked analyses for scenario-based hierarchies”, which also builds an emulator to predict future climate from arbitrary CO2 forcing scenarios.

Thank you, we will include this relevant citation in our revision.

Technical corrections:
We will address each of these valid points in the revision.
1. Page 3350, line 26, “Parametric error” is probably better described as “parametric uncertainty” as the parameters may not have an operationally defined meaning. There is also no mention of simulator discrepancy in the introduction, which can dominate over parametric uncertainty.
2. There is possibly some confusion in terminology in section 4.1 over the EOF de- composition. I had thought that the name “EOF" was interchangeable with “principal component" (or "loadings") and represented the spatial patterns observed in the data. What the authors call the principal components (the right singular vectors in their notation), I would have called the scores.
3. Page 3361, line 7. “D is the 564 × 564 diagonal matrix of eigenvalues". Should this be, “D is the 564 × 564 diagonal matrix of the square root of the eigenvalues of the covariance matrix Y Y T "? Note also that this only applies in the case that Y has been centred.
4. Should the title of section 6 be “Emulation of the effect of the representative concentration pathways" or something similar