We are grateful for these detailed, constructive and very useful comments. Our responses are in italics.

Anonymous Referee #2

In this paper the authors present a new climate model and introduce a new emulator on both the parameter and scenario space, which extends previous work and emulates multiple timepoints. Using a principal component decomposition, they develop independent emulators for every PC. The efficiency of the emulator is tested with goodness of fit measures, as well as cross validation, and the climate model is also compared with observations.
This work has elements of interest in developing and extending some statistical methodology to multiple times, but it is not clear how the development of a new climate models has helped in the emulation problem, given that the main topic of the journal is climate models, not statistical methodologies. Also, this technique has some important drawbacks that were not reported and should deserve discussion. To be suitable for publication in this journal, this paper needs major revision.

General comments
• Emulation is a statistical procedure that is independent of the particular climate model that one wants to emulate. Most of this paper is dedicated to the emulator, and in the introduction this is presented as the main contribution. The part that describes the climate model seems in this work only functional to the emulator rather than having an interests on its own. The topic of this paper seems not within the range of this journal, whose focus is in climate model themselves. I would suggest to at least redo the introduction redefining the objectives and trying to stress more the importance of this work in the context of climate model development.

Yes, this is a good point. Here, the "climate model" under development is in fact the emulator. The purpose of this new "statistical climate model" is to replace the underlying physical climate model in situations where computational demands require it. The emulator has already been applied to a range of such applications, applications that would not have been possible using a simulator of this complexity. We will clarify this motivation for the work, and also will discuss and cite the applications to demonstrate that the emulator is a useful climate model in its own right.

A number of the referee's comments below appear to come from the assumption that this is intended to be a statistical theory paper. Rather we believe it is better viewed as a model development paper, as befits the journal, and hence represents a description of the choices that were made in the development. We agree that the methodology could be extended further to emulate more complex climate models, different time domains, different applications, and we also agree that changes to the methodology (e.g. Gaussian processes) might be required to achieve some of these objectives.

• The description of the statistical model could be improved in terms of mathematical rigor and clarity, there are no clear equations describing what are the assumptions, and a discussion on their validity. For example, at the beginning of Section 4.2, the authors write: "For each PC emulation we build a linear model from all 28 parameters, and then allowed the stepwise addition of ten quadratic and cross terms". This is the main emulator, but is not clear how you added the terms, why you added only ten of them, and why did you pick that particular subset.

We will improve the rigour of our mathematical description. In the specific example given however, the description is sufficient to replicate the procedure followed. It hinges upon the use of stepAIC, a well-known tool (which is fully cited here) for stepwise selection of model terms subject to the constraints of information criteria.

As noted in the manuscript, one or more of the ten terms was removed by stepAIC in approximately half of the emulators when the BIC criterion was applied to allow model shrinkage. This suggests that the choice to allow ten terms was not unreasonable. More rigorous approaches to determining the number of terms are indeed possible. The most rigorous is perhaps to progressively add terms individually and test the resulting model under cross-validation until the improvements plateau. However, the benefit of such an approach is likely minimal, and the computational effort required highly demanding, especially given that each separate climate model output is associated with ten emulators. We will, however, apply this approach on at least a subset of emulators in the revision in order to address this question.

• Where was it shown that "parametric error" (which would be perhaps better referred as "parametric uncertainty") is significantly larger than the "code error" (again, "code uncertainty" sounds better) in this study? My guess is that given the ambitious goal of emulating in both parameter and scenario space and given the relatively modest ensemble size, the parameter variability will be very large so this could be justified here. However, this will not be true in a low dimensional calibration problem, where GP emulation is the more suitable approach. So the input dimension of this study (28 in this case), along with a sparse parameter design, justifies a simple statistical model. This should be pointed out as it helps the reader to understand why the emulator does not rely on a more standard approach.

Thank you for the suggestion. We will discuss the relative merits of GP and parametric emulation in the revision.

• The EOF decomposition is based on the assumption that a small number of principal components is able to reproduce the spatial structure of the data, but I don't find this obvious. It is true that the principal components are able to explain the variance of the data, but this does not necessarily imply that that they can also describe the spatial correlation. Closer data in space are expected to be more correlated, and the modeling of their dependence is a wide topic whose review is beyond the scopes of this work. Nevertheless, depending on the nature

of the correlation, low rank approximation could not be sufficient to reproduce the statistical features of a spatial field. Some discussion in this regard is needed.

We agree that we have not considered pathological cases in which important modes of spatio-temporal covariance in the data are not spanned by the mean of the data plus high-order modes of variance of the ensemble. We will discuss this in the revision and we will extend the analysis to demonstrate the degree to which the first ten components are able to reproduce the spatial structure of the data.

• The use of Chebyshev polynomial (3) and (4) implies some limitations in scenario emulations that were not discussed.
– This approach allows emulation for only a fixed number of years. More specifically, if one wants to emulate over a specific number of years, he can only use the model output for those years over which the prediction is sought. This was not an issue here, since the ensemble was generated with the same endpoint, but it could severely limit its application to other ensembles where different runs have different endpoints (e.g. some CMIP5 runs end in 2100, some in 2300). The current statistical model will force the user to simply discard all the simulated years beyond the timeframe of interest, possibly losing valuable information. Besides, if one wants to use the emulator to predict at a further time point, he would need to refit the statistical model. This implies SVD of moderately large matrices and to fit a model for every EOF, and that could be computationally demanding. Since the authors have used only decadal averages and a very coarse spatial resolution, I guess they can overcome this limitation with sufficient computational resources, but for finer time scales (years, or months) or with a state-of the-art resolution this will be not trivial. To reduce the size of the spatial field, a different approach could be to consider independent emulators over some predefined regions. This will ignore the dependence between regions but it will be scalable. Also, the fit will be even harder when using statistical models which account for dependence across parameters (e.g. GP emulation). A paragraph discussing the statistical model complexity and its scalability is needed.

*Yes, we will discuss scalability issues. It is perhaps worth noting here that we have applied the outputs of this emulator at *daily* timescales. To achieve this, the decadally averaged seasonal climate fields are interpolated into monthly annual climate fields. These are provided to ClimGen, which adds artificial daily weather to generate an output suitable for crop simulation. This discussion is clearly outside of the scope of this paper, but is mentioned here to demonstrate that the approach is far more flexible that it may at first appear (and represents a substantial advance upon pattern scaling which requires similar procedures to convert a single climate pattern to daily "weather".)*

– This method violates causality. The decadal temperature for, say, 2055-2065 is emulated by also using the values of the forcing scenario after that date. This can be fixed by doing independent emulation of every decade, but I don't see any easy modification of this method to generate a time series which avoids this.

Whilst this is technically true, and an excellent point, it is not a practical problem. The emulator, as any statistical model, is not constrained by causality, nor indeed is it constrained by any physical law. However, the emulator has been trained upon data produced by the simulator, which is governed by physical laws. Thus, while the first chebyshev coefficient TC1 is a function of 2100 CO2 (apparently violating causality as the 2050 climate is a function of TC1), the climate at 2050 is also a function of TC2 (et al), and TC2 is uniquely defined by 2050 CO2 and TC1. Through such complex dependencies (TC1-3, EOF1-10, 10 time slices per EOF), the statistical fit is able to reproduce simulated climate at 2050 and be ***approximately*** *independent of the 2100 CO2. We will discuss and expand upon these issues in the revision.*

– With this parametrization it is not possible to emulate scenarios with abrupt jumps or drops. This is of course not an issue for impact assessment, but it could be if one were to use emulation with a different goal in mind, such as understanding the physics of the system by increasing the signal to noise ratio.

We partly agree, although we note that abrupt change can, in theory at least, be represented within the EOFs themselves. An EOF is not required to exhibit the same spatial pattern throughout its temporal history, as has been discussed in reference to Fig 1b, and so would reflect any abrupt change that contributed significantly to the ensemble variance. However, although the output is not required to be smooth in this sense, we agree that the assumption of smooth input can be a restriction and this needs discussion in the revision.

• I found Section 4.4 quite unclear. My understanding is that the EOF were obtained via SVD of Y, so they represent the decomposition of the whole space/time field for all the runs (or a subset of them for crossvalidation). What does it mean that here the EOFs are averaged over space at each time slice? I would like to have more details in this section before I can further comment.

The reviewer is correct, we will clarify this in the revision. Although each EOF represents a component of the variation of the ensemble across time, space and between ensemble members, each EOF, like each ensemble member, is a (3D) spatio-temporal field and can be averaged in space to give a function only of time.

Specific comments
• p.3350 l.15. The results in this paper are of two types: simulator results and emulator results. The former are compared with empirical data while the latter with other higher-complexity models. It is worth pointing out this difference when mentioning validation.
Agreed
• p.3351 l.10. What needs to be emulated here is not only a "high dimensional output", but an output whose dimensional indexes are space and time and are therefore physically meaningful. This was not acknowledged throughout this work and the resulting statistical methods violate causality in time and might not be the best approach for spatial dependent output as well (see general comments).
We will discuss this.

• p.3351 l.25-28. The emulator is a simple statistical model, especially in this work where only linear models were used, so it is possible to have gradients in a simple form. However, it is in principle possible to compute gradients directly from the primitive equations underlying the climate model. Besides, this feature of the emulator was not used here, so there should be some references on works where this was.

This is an interesting comparison and will be discussed in the revision.

• p.3351 l.28. The word "calibrated" is correct here, but should be augmented with references to the MPEF ensemble on p.3357-3358, since the reader has no information on the ensemble at this point.

Agreed

• p.3352 l.11. What is the meaning of "self-consistent" here?

Revised text will clarify.

• p.3353 l.15-17. Section 5 compares the HDDs and the CDDs from the climate model with observational data, it is not using any emulator. Also, Section 6 is validating the emulator with the CMIP3 and CMIP5 ensembles, not the actual climate model.

Agreed, and will be clarified. Our intention was to validate the simulator and DD methodology with respect to observational data. For future change we validate the emulator against the CMIP ensembles rather than the simulator itself. Our rationale here is to demonstrate that the emulator provides meaningful predictions, not just to demonstrate that it can reproduce the simulator. In a sense, we are validating both the emulator and the simulator in this step.

• p.3356 l.26. It is perhaps more appropriate not to refer to climate runs with different physical parameters as "realizations", but simply as "runs". Realizations imply independence across the different values in the parameter space. Although this might be true for an initial condition ensemble, this is not the case in a perturbed physics experiment.

We are happy to make this change if it is clearer.

• p.3357 l.22. Specifying that the key model outputs are five helps the reading in line 27

Agreed

.• p.3357 l.23-26. This part could benefit from more details, i.e. how many parameters were eventually used, is there any evidence that some cross-terms are more significant than others, etc..

Agreed

• p.3358 l.4. Why does the word end looks slanted?

This was intended to emphasise that all five plausibility tests must be passed. Although grammatically not needed, we have found that this point was easily missed. We are happy to remove the italics.

• p.3358 l.1. A reasonable model state is meant with respect to preindustrial conditions here, since the plausibility test on line 19 has a higher temperature range, is that correct?

Yes, we will clarify.

• p.3358 l.23-25. On lines 6-7 you mentioned that only 10 of the 500 parameter set where classified as plausible, yet here there are 188 simulations which pass the modern day plausibility test. I don't understand the transition from the MPEF to the MPSF set.

We will clarify. Only 10 parameter sets in the MLH ensemble were plausible. This is the Latin Hypercube training ensemble, as distinct from the emulator filtered MPEF or simulator filtered MPSF ensembles.

• p.3358 l.22. Why are the energy balance bounds larger than on line 1?

During the emulator filtering process we apply tighter filtering bounds. The emulator is cheap to run and we can afford this additional computational expense. This is done because we know that the emulators are imperfect and so this step will reduce wastage during the simulation filtering. It also has the advantage of producing more simulations near the center of the prior, the region that is most plausible. We have discussed this in previous publications, but agree that this should be clarified here also.

• p.3359 l.17. Have you tested how much the results of Section 6 would change with more Chebyshev polynomial? Or with another functional basis?

No. This would be highly demanding as it would require a simulation ensemble (~3 CPU years) for each alternative profile considered.

• p.3360 l.6. The Maxmin Latin Hypercube is a particular design for the (A1e, A2e, A3e, A1, A2, A3) which maximizes the distance across the parameters in the design. In this case however, we are interested in the functions resulting from these parameters, not in the parameter themselves. Does this design result in some particular property for the functions as well? Besides, what would be the meaning of maximizing the distance across parameters that generates two different inputs in the model (CO2 and CO2e)?

We believe it is correct to assume the six parameters are (approximately) orthogonal in the design. Non-orthogonal inputs can lead to very unreliable fits as it unclear which parameter is driving the change in the emulated output.

• p.3360 l.21-22. A reference on Section 5 might help here.

Will include

• p.3361 l.28. What to you mean by "DJF SAT variability"? You mean the variance across the three months for daily temperature? Or the variance for monthly temperature? If this is the case, why not calling it simply "variance"? It would make the reading easier.

Will clarify

• p. 3361 l.23. The citation to the R development team should be for the year 2013.

Will correct

• p. 3363 l.10-12. PC components are linear combination of the model output, and it is generally hard to give them an interpretation in terms of the original data. What kind of physical processes are reflected by linear combinations of high order PCs?

We will discuss further.

• p.3363 and Table 2. How was the fitted R2 computed? Was it computed independently for all the simulations (i.e. across all rows of the Y matrix in (5)), and then averaged out across the 564 elements of the ensemble? Also, my understanding is that for each PC a different emulator was build and an R2 was computed, but shouldn't it be more of interest to see the incremental contribution of the different PCs? In other terms, for the kth emulated PC pne could consider the field reconstructed with the 1, . . . , k PCs. This should presumably have a increasing R2.

We will provide this analysis

• p.3363 l.20. I don't fully understand this discussion about adding orthogonal terms. It is true that every PC adds predictive power, but this is true when you consider the joint contribution of all PCs, which you haven't done here (and even in this case, the contribution can be marginal anyway). Do you use the term "emulator" to indicate the reproduction of the space/time fields from the single PC or all of them?

We agree, and we will discuss this in the revision. We will expand the analysis to consider the joint contribution of all PCs, progressively adding components, as the referee suggests.

• p.3364 l.2-5. Reproducing EOF might be a necessary condition to approximate the space/time output, but it is not obvious why this is also a sufficient condition (see general comments). A more interpretable summary could be the spatial or temporal variance of the emulator vs the simulator, or a comparison of the temporal autocorrelation (assuming maybe a simple AR(1) model after some de-trending). It is hard to understand what does of this underestimation of the PC variability implies in terms of the lack of characterization of the space/time properties of the field.

We will extend our analysis to consider how well the fields are emulated, not just the PC scores as at present.

• p.3366 l.9. Remove "as". • p.3368 l.3 "CO2 is only an input". Without the comma. • p.3368 l.8. The future transient period should be 2000 to 2100 AD. • p.3369 l.5. What is the "CW05" fit?

Will correct

• p.3369 l.18-19. This paper has shown that a model with very coarse spatial and temporal resolution can be emulated and has shown better results than pattern scaling, but pattern scaling does not need large matrix computations. If we consider for example the RCP experiments for the CMIP5 CCSM4 ensemble, L and R will have 192 × 288 × 251 rows and 24 columns, which means that in double precision each matrix will be approximately 2.5 Gb. It will not be trivial to store these matrices in the RAM, and make scalable computations, and it will be even harder in the future with higher resolution models or for statistical models which account for parameter dependence. I am not suggesting that this method is not a good first step, but scalability in the context of climate models should always be taken into account.

We will discuss.

• Fig.1. For subplot a, why there are no labels on the x and y axis? For subplots b-f, the resolution of the model is 64 × 32, as stated in p.3352 l.24 and in p.3360 l.25, so why is there a 36 × 18 grid here? The data look constant over some grid, but that is the one that is represented. I would suggest plotting the original T21 grid. Also the reported data min and max are too small to be readable, larger fonts are needed. Why they are reported only for subplots e-f? What does the label "data 0" on subplots c-f means? Subplot b has fonts which look smaller than the others.

We will redraw

• Fig.2. A reference on the names on this table could help (I know they come from SIAM-WORLD, but refer to a certain page or table on a literature reference). Also the underscores in the x labels should be deleted, and there must be a labeling on the y axis. The figure is too small compared to the size of the caption, can it be made larger?

We will redraw

• Fig.3. Every row has the same x – y scale, and this might hide some of the polynomial misfit for RCP 4.5, 6.0 and 8.5. I would suggest using the log scale. Since the temporal scale is the same for all columns it is not necessary to repeat the labels for every row, this redundancy looks unnecessary. Larger fonts for the labels and legend might help, too.

We will consider redrawing, although this was deliberate. Differences will of course look more significant when the scale is redrawn, but the fixed scale puts the errors in context. For instance a few ppm error might look significant in RCP2.6 when viewed in isolation, but it is insignificant when compared to the range of emissions that arise under different possible future forcing scenarios.

• Fig.4. See comments for Fig.1.

We will redraw