Final author comments for "Large ensemble modeling of last deglacial retreat of the West Antarctic Ice Sheet: Comparison of simple and advanced statistical techniques", by D. Pollard, W.Chang, M.Haran, P. Applegate and R. DeConto.

We thank the reviewers for their careful and helpful comments on the original version of the manuscript. Our responses and planned changes are described below point by point, with reviewer text in italics. In summary, the main planned changes are:

- All "future" simulation segments in text and figures will be removed (Reviewer 1).
- References to upcoming work and papers will be reduced, and we will make clear that this paper stands on its own (Reviewer 1).
- Alternate "close-to-Gaussian" approach to misfits and scoring will be added, discussed, and results compared, in sections 2.3 and 2.4, Appendices B and new C (Reviewer 2).
- Spans of model results over all runs will be shown to encompass the various types of observations, in new Appendix D (Reviewer 2).
- Description and discussion of the advanced statistical techniques will be expanded, and their role will be made clearer, in sections 2.5 and 5 (Reviewer 2).

Reviewer 1:

Overview:

The submitted paper presents results from a large ensemble of ice-sheet model simulations of the West Antarctic Ice Sheet through the last glacial termination and into the future. The ensemble aims to explore a broad envelope of parameter space, and two different techniques are employed to assess the results. As far as I can tell, the primary justification for the paper lies in the intercomparison of so-called 'simple' and 'advanced' statistical techniques, rather than the presentation of realistic simulations of the deglacial and future states of the ice sheet.

This is correct: the primary purpose of the paper is to compare 'simple' and 'advanced' techniques (see next response below).

Overall the paper is well-written and clearly laid out, with thorough explanation of the salient aspects of the study and sufficient reference to the preceding studies on which it builds. The figures are clear and effective. As a methodological paper it is clearly

well-suited to GMD.

General issues:

I have detailed a few points lower down that I think need further explanation or clarification, but I have two more general issues with the manuscript as it stands.

Firstly, there are numerous (at least 8) instances in the text (p6 lines 22/23; p7 lines 18/19; p13 lines 19/20; p14 lines 8/9; p16 lines 1/2; p16 lines 17-21; p18 lines 6-8; p18 lines 23-25) where the authors refer to 'future work' that will either develop or change some aspects of the study as presented here. Whilst it is of course quite usual that submitted work forms part of a project that is ongoing, I found the repetition of these statements quite off-putting in the sense that they give the reader the impression that the current study is in some way 'incomplete', or worse still, inferior with respect to something similar that is being prepared for another journal (for example, the reference to Pollard et al., 2015b, which is a paper that is only 'in preparation'). I think the paper should be able to stand alone, and if important aspects of the study are either yet to be developed, or modified, then what is the rush to publish seemingly incomplete work here? Will the forthcoming papers build on this one, or undermine it?

As mentioned above, the primary purpose of this paper is to compare the simple and advanced techniques, using the same large ensemble (stated in section 1). We agree that the numerous instances referring to other work may detract from this purpose. The current paper definitely stands on its own and is complete, and the results do not depend on or will be changed by any of these instances. Accordingly, we will (i) emphasize more the purpose of this paper in the introduction, (ii) remove many of these instances where they do not contribute to the M/S, (iii) where the follow-on paper (Pollard et al., 2016) is first mentioned, explain that it deals solely with specific glaciological aspects, not statistical, and does not undermine or alter the results here at all.

In the concluding Section 5, we plan to still mention several avenues and plans for further work, which all concern glaciological aspects, not statistical. As the reviewer mentions, this type of discussion is quite usual in the concluding sections of papers.

The second issue I have with the manuscript as it stands is the inclusion of the 'future' scenario modelling. The title and majority of the paper deal with the deglacial, and since the primary purpose of the paper is to compare results from different statistical

methods (for which any results would do) I see no reason to include the additional 5000 year experiments. They are barely discussed in the paper and have no relation to the deglacial experiments. Furthermore, as detailed below the basis for the 6C/2C air/ocean warmings is not clear. If they are arbitrary, then what is the justification for adding them to the end of a supposedly 'realistic' deglacial run? And if they are meant to represent a future emissions scenario such as RCP 8.5, then some explanation is needed to clarify why this is used rather than, for example, RCP 6 or any of the others. To my mind it looks like these data have been added to the paper somewhat opportunistically, rather than for any particular purpose. And by the authors own admission these simulations use a climate warming that is 'very simple' (p14, line 7), and the future simulations themselves will be presented in more detail in, once again, the forthcoming Pollard et al 2015b paper currently 'in preparation'. On this basis I think these arbitrary extensions to 5000 CE should be removed and saved for the other pending publications.

We will remove all mention of the simple "future" extensions in the M/S. These extensions were part of earlier work exploring the response to future warming, but have been superceded by further work with more realistic future climate RCP scenarios (with references cited here). This is a natural extension of the past simulations here, but we agree that they do not add to the purpose of this M/S (and again, do not change the statistical results at all), and belong appropriately in subsequent papers.

Specific points:

p6 - I think the justification for not using the 'drastic ice-retreat mechanisms' of Pollard et al 2015a should be more fully discussed. Either these mechanisms are necessary for realistic simulations (as argued in the EPSL paper), or not. Or do the processes only happen during warm periods and not cold periods? It seems that any complex statistical analysis of results is only useful if it helps reduce uncertainties, but if the largest uncertainty is ignored (ie uncertainty over the inclusion or exclusion of 'drastic' mechanisms) then the results are inherently biased. It would be useful to see how the results change when the 'drastic' mechanisms are included.

These mechanisms are only triggered in warmer climates than present, as the reviewer suggests. They do not play any roles in the glacial-to-deglacial sequence of the last ~40 kyrs, as confirmed by tests (not shown here). We will note this in the model description section 2.1.

p7 - Liu et al 2009 present a transient run that ends at 14 ka BP, so what is used to drive the model from 14 ka to present?

Although the Liu et al. (2009) paper only describes results to 14 ka BP, their simulation has been extended to the present, which they call the "TraCE-21k" experiment; see www.cgd.ucar.edu/ccr/TraCE. We will note this in the references and acknowledgements.

p7 - what is the basis of the 6 and 2 C air / ocean temp increases? RCP 8.5 after 150 yrs equals c. 6 C air temp above present, but CMIP models suggest 6 C air would equate to 1.5 C in the ocean, not 2 C, which presumably could affect the results presented here? Similarly, the extended RCP scenarios define warming trajectories that increase steadily to 2300, and remain constant thereafter, rather than flat-lining at 2150 as implied here.

This is no longer pertinent since all text and figures concerning the future extensions will be removed (see above).

p7 - since these "future" simulations are regarded as unrealistic, why include them? Particularly if the 'drastic ice-retreat mechanisms' aren't included.

As above, no longer pertinent.

p15 - 'Macintosh' should be 'Mackintosh'

We will correct this.

Fig. 5 - y-axis label is 'sea level rise (m)', which implies that it is showing time-varying rates of change in sea level, but I think it is actually showing the change relative to present? Otherwise the value of c. -6 m from -20 ka to -15 ka could be read as indicating that the sea level was falling constantly by 6 m through that period.

We will change the label to “equivalent sea level (m)”.

Reviewer 2:

The submission can be an informative (and relatively succinct) comparison of two different approaches to making inferences about past ice sheet evolution given modelling and paleo observations. Some specific issues (including some mis-citations) are detailed below. There are four key deficiencies that have to be remedied (to change the above "can be" to "is"):

1) Currently there are no plots nor discussion of model fits to constraint data and as such it is not clear whether this ensemble actually covers the constraint data.

We will add a new Appendix D with extensive figures and some discussion, showing the span of results of all 625 runs of the large ensemble (LE) compared to observations, for the various past data types. This will consist of individual plots for specific sites for Relative Sea Level and cosmogenic elevation-age data, and as a single plot for modern uplift rate sites. Also, maps of grounded-ice probability computed from the LE will be compared with maps of reconstructed grounding line positions at specific past times, and similarly for grounding-line distances vs. time along paleo-troughs of the major embayments. These plots (already made) will show that the span of model results does by and large encompass the observations with no serious outliers, as required for meaningful interpretation of the statistical LE results.

2) The handling of data uncertainties for all the misfit metrics needs to be spelled out (some treatments are spelled out, but not all). Eg, TROUGH will have dating and downscaling/resolution uncertainties. If these uncertainties are ignored, the inferences based on these metrics are biased and incorrect.

Considerably more detailed description and formulae of all misfit calculations will be given in an expanded Appendix B. These will aim to give a complete description of all calculations.

3) how are data weighted within each class? If no weighting is done, then the statistical

modelling is assuming all data/model residuals are not correlated, which is incorrect (though commonly implemented...).

Within each class, intra-data-type-weighting is done, very much the same as in Briggs at Tarasov (2013), for past data with individual sites: Relative Sea Level, elevation-age, and uplift rates. Full details will be given in Appendix B.

4) There has to be justification for giving all data classes the same weight. There are only 8 RSL data sites, all located on the periphery of the ice sheet. There is no basis to give this geographically restricted data the same weight as, for instance, the RMS error between the dynamically modelled and observed present day ice sheet.

We agree that this is a significant issue, but take a different strategy than in the Briggs et al. papers. Here, we assume that each data type is of equal importance to the overall score, and that if any one individual score is very bad ($S_i \approx 0$), the overall score S should also be ≈ 0. This corresponds to the notion that if any single data type is completely mismatched, the run should be rejected as unrealistic, regardless of the fit to the other data types. The fits to past data, even if more uncertain and sparser than modern, seem equally important to the goal of obtaining the best calibration for future applications with very large departures from modern conditions. This differs from the "inter-data-type" weighting based on "volumes of influence" in Briggs et al., which is interesting and logical, but we suggest is heuristic and not the only reasonable way. Also see the response to "Gaussian forms" point (4th below).

If the "advanced statistical method" does use a complete error model that addresses points 2-4 above, then this should be made clear in detail. Ie, are you saying that we can ignore all these issues, do simple latin hypercube sampling (albeit with a large enough sample, but still orders of magnitude smaller than required for proper MCMC), and get roughly the same result as a complete Bayesian calibration determination of the posterior (ie with a complete error/uncertainty model that accounts for uncertainties in the constraint data, structural uncertainties, and correlation between residuals and that covers the constraint data set)? If so, then this claim need to be much more clearly spelled out.

We acknowledge that some sentences in the M/S were somewhat unclear regarding this point, which will be clarified. In this paper, the advanced techniques do not use a Latin HyperCube large ensemble (LE), but are applied to the same LE as the simple averaging method, which is a 625-member LE with full factorial sampling. The purpose of this paper is just to compare

statistical results of the two methods, with the advanced techniques acting as a benchmark. In previous studies (Applegate et al., 2012; Chang et al., 2014), the advanced techniques yielded successful results when applied to some relatively small-sized LE's with coarse Latin HyperCube sampling, for which the simple methods failed. This is because the interpolation capability of the advanced techniques (emulation, MCMC) is much better than the simple method (essentially none). However, this distinction depends on the size of the LE and the coarseness of the sampling; somewhat larger LE's with Latin HyperCube sampling and fewer parameters can be amenable to the simple method. This will be briefly noted in the conclusions, where we will emphasize that it is not otherwise addressed in the paper.

Once these (and the comments below) are addressed, I would agree with Nick Golledge as this being a methodological paper that is well-suited to GMD.

 # Specific comments:

 # How is relative sealevel computed? What visco-elastic earth model is used and is geoidal deformation computed?

The bedrock response component in the ice sheet model is a basic ELRA (Elastic Lithosphere Relaxing Asthenosphere) model. Sea level vs. time in the ice model itself is prescribed from ICE-5G. These are noted in the model description section 2.1.

The calculation of relative sea level at specific grid points for comparison with RSL geologic data is as in Briggs and Tarasov (2013), and will be described fully in Appendix B.

"Tarasov et al. (2012) used Artificial Neural Nets in North American ice-sheet modeling to fill in parameter space between LE simulations, and have mentioned their potential application to Antarctica (Briggs and Tarasov, 5 2013)."

 # actually this was as much if not more of a "calibration" as the authors' "advanced statistical technique" and should be clearly stated as such. That 2012 paper also used MCMC to compute a posterior distribution of ensemble parameters given fits to paleo constraint data. The reason that "calibration" wasn't used in the title of that paper was 1) ensemble didn't cover data constraints (attaining coverage is a big challenge given the large size of the constraint data set), and 2) it had an incomplete error model especially with respect to quantifying structural uncertainties. Unfortunately, "Calibration" has become a poorly understood buzzword whose meaning is being watered down in

some recent ice sheet relevant publications. To me, if "calibration" is not confidently estimating the probability distribution and thereby the uncertainties of predictions (with the unavoidable clear specification of uncertainties not accounted for), then it should not be called calibration. But this may be a loosing battle...

We will rephrase the relevant sentence in order to (hopefully) resolve this concern, as follows:

> Tarasov et al. (2012) used Artificial Neural Nets in their LE calibration study of North American ice sheets, and have mentioned their potential application to Antarctica (Briggs and Tarasov, 2013).

"Then the geometric (logarithmic) average of the 8 individual Si 's is taken to yield the aggregate score S for each run"

*# This choice makes no sense to me and needs to be justified. RMSE is effectively log(Gaussian). So your weighted score is (logGauss1*logGauss2*..)^1/8. How does one interpret this? If you are using a non-Gaussian error model, then what is it?*

We propose that the formulae chosen for misfits and scoring are somewhat heuristic and there is more than one reasonable approach, and that strict adherence to Gaussian error model forms is not the only possibility. In section 2.3 we will add the following text to explain and justify this viewpoint:

> One approach to calculating misfits and scores is to borrow from Gaussian error distribution concepts, i.e., individual misfits M of the form $[(mod\text{-}obs) / \sigma]^2$ and overall scores of the form $e^{-\underline{M}/s}$, where *mod* is a model quantity, *obs* is a corresponding observation, σ is an observational or scaling uncertainty, $\underline{M}$ is an average of individual misfits over data sites and types of measurements, and *s* is another scaling value (Briggs and Tarasov, 2013; Briggs et al., 2014). However, the choice of these forms is somewhat heuristic, and different choices are also appropriate for complex model-data comparisons with widespread data points, very different types of data, and with many model-data error types not being strictly Gaussian. In order to determine the influence of these choices on the results, we compare two approaches: (a) with formulae adhering closely to Gaussian forms throughout, and (b) with some non-Gaussian aspects attempting to provide more straightforward and interpretable scalings between different data types. Both approaches are described fully below (next section, and Appendix B). They yield very similar results, with no significant differences between the two, as shown in Appendix C. The second more heuristic approach (b) is used for results in the main paper.

Accordingly, we will make a significant addition to the paper, adding a new set of formulae for misfits and scores, that do adhere closely to Gaussian error forms. We will call this "approach (a)", vs. "approach (b)" for the existing set of formulae in the M/S. Both sets of formulae will be described in an expanded Appendix B and in Section 2.4. Comparisons of all results will be shown for both approaches in a new Appendix C, which will show no significant differences (figures already made), indicating that they are robust and independent on the choice of approaches to misfits and scoring.

"It differs from from the weighting in Briggs and Tarasov (2013) (their "inter-data-type"), which is algebraic and depends heavily (80%) on the fit to modern ice distribution."

 # This is incorrect. The weightings are for the RSME score components, but the final weighting is e to the power of the sum of these normalized components (ie assumes a pseudo-Gaussian error model). This is therefore not algebraic. Furthermore, Briggs, Pollard, and Tarasov (2014) should be cited instead. They give a corrected inter-datatype relative weighting of < 50% for present-day data (Coauthors should know the papers their names are on, rap knuckles.., :)).

The relevant sentence in Section 2.4 will be rephrased, avoiding specific values:

> Of the two approaches, this most closely follows Briggs and Tarasov (2013) and Briggs et al. (2014), except for their inter-data-type weighting, which assigns very different weights to the individual types based on spatial and temporal volumes of influence (Briggs and Tarasov, 2013, their sec. 4.3.2; Briggs et al., 2014, their sec. 2.2).

"3. Consistent with trends in recent Antarctic modeling studies (Ritz et al., 2001; Huy20 brechts, 2002; Philippon et al., 2006; Briggs et al., 2013, 2014; Whitehouse et al., 2012a, b; Golledge et al., 2012, 2013, 2014), the greater total Antarctic ice amount at the Last Glacial Maximum is less than in earlier papers, equivalent to 5 to 10m of global equivalent sea level below modern"

 # Incorrect citation of Briggs et al, 2014: Their confidence interval for LGM Antarctic ice volume excess has an upper bound of 14.3 m eustatic equivalent, with lower confidence is > 10 m, and one of their single best fit runs has an excess of 13.2 m. Furthermore, they raise the point that their (well our) model had insufficient grounding line

response compared to proxy paleo data, suggesting that LGM grounded ice volume could be under-estimated. So there is no basis to lump this in with other studies claiming </= 10 m of eustatic sealevel equivalent.

This is a valid point; we will accommodate it by changing the range "5 to 10 m" in the sentence to "5 to 15 m". This is still less than the > 20 m in older papers (such as CLIMAP, Denton and Hughes, 1981).

"For ELEV: the minimum squared mismatch of ice elevation and time, within the constraints of descending elevation trend, each relative to the observational uncertainties of elevation and time"

#Bit unclear. Is this the same error model as Briggs and Tarasov 2013?

Yes, very close to the same. Full details will be given in the new Appendix B.

A. Kergweg:

Dear authors,
In my role as executive editor I ask you to move the Code Availability Section to its usually place after the conclusion but in front of the Appendix when revising your article.
Thanks, Astrid Kerkweg

We will move this section as requested.