Response to comments by Anonymous Referee #1 on
"How well do different tracers constrain the firn diffusivity profile?" by
C. M. Trudinger, I. G. Enting, P. J. Rayner, D. M. Etheridge, C. Buizert, M. Rubino, P. B. Krummel, and T. Blunier, doi:10.5194/acpd-12-17773-2012

RC: General comments:
The article presents a new method for the calculation of best estimate firn diffusivity profiles and associated uncertainties. Firn diffusivity is an essential parameter for the modelling of trace gas transport in firn. In my opinion, this study is thus important for scientific issues related to trace gas age distributions and inferring atmospheric trace gas trends from firn and ice core data. I think that the manuscript would be read and used by a larger audience if it was written in a shorter and more goal-oriented way. In its current state, I find the manuscript's main conclusions weak and many contradictory intermediate conclusions in terms of which tracer(s) is (are) best to constrain diffusivity render it difficult to follow. However, I think that the methodology used has a clear potential for providing recommendations for future firn air pumping operations. Important issues might easily be addressed through similar model tests than those performed such as which depth resolution should be used? How results are affected by calibration biases? What is the effect of seasonality related firn layering? Such questions are implicitly raised in the model design but not specifically tested in order to be really answered.
AC: We thank the reviewer for their thorough review. We have tried to rewrite the paper so that it is shorter and more goal-oriented, and to bring out the main conclusions more clearly. We have deleted some information, moved other information to the Supplementary Material, and tightened up some sections.
Our method has many possible applications, and it is only possible to cover some of these in one paper. In some cases we have added extra cases as suggested by the reviewer, but it has not been possible to add all extra calculations that were suggested.

RC: I also have some major concerns that are detailed below:
- about the definition of the equifinality criterion used (see comments 2-4, 9 and 27) and other issues related to the model design (17, 26, 31, 32, 35, 40, 57, 64)
AC: These points are addressed individually below.

RC: - about the important portion of the manuscript devoted to synthetic cases A and B (_8 pages of text, one complex table and 6 figures). The complexity of the design and analysis of these scenarios, and the fact that the ranking of tracer efficiency to constrain diffusivity is obviously depth and real word uncertainty dependent render their usefulness limited in my view. In contrast, DE08-2 results (section 3.4) are presented in only 10 lines and South Pole results (2 boreholes) in 12 lines. A more in depth analysis of model results obtained with "real" data is needed in my view.
AC: Our reason for considering the synthetic cases is that we wanted to use them to establish our methodology and demonstrate the principles for comparing the value of different tracers for cases where we know the exact characteristics of the noise and the "true" diffusivity and concentrations. Further justification for the synthetic cases is given below at comment 21. We have not removed the synthetic cases, but have condensed their discussion to focus on the most important conclusions, and have moved the Synthetic B figure and table and some description to the Supplementary Material.

RC: - about the lack of clarity and clear conclusions in many parts of the manuscript, which renders it difficult to read.
AC: We have tried to improve the clarity, and focus on providing clear conclusions.

RC: I make many suggestions below aiming at improving the analysis of the results and reaching clearer conclusions. I hope that they will help the authors to improve this article which I think has the potential to become a very valuable contribution to the improvement of firn modelling.

Specific comments:
1 - p17774 l20-23: This sentence implicitly suggests that previous studies over the last two decades paid no attention to uncertainties. It provides a recommendation that I find difficult to implement: all uncertainties (e.g. non modelled processes) are not easily quantified precisely. Please reformulate.
AC: Our intention was certainly not to suggest that previous studies, including our own, have paid no attention to uncertainties. We have replaced "allow" with "assist in" to avoid this interpretation.
We don't believe that the recommendation to use multiple model configurations with multiple parameter sets rather than a single model configuration with the best parameter set is particularly difficult. When diffusivity is calibrated, often a number of alternative parameter sets with adequate fit to observations are found, either before reaching the 'best' solution, or with different model/optimisation configurations or process representations. We are not suggesting the use of multiple models here.
Of course it is not easy to quantify all uncertainties precisely, but our approach gives a starting point, led by the latest information on what processes are believed to be uncertain. When there are different possibilities for uncertain processes (dispersion in the lock-in zone is a current example), unless or until some options have been shown to be unlikely, it is valuable to include different model configurations that represent different possibilities, in calculations to infer atmospheric records, to quantify how much these uncertain process affect the results.

2 - p17775 l5-13: This sentence describes physical processes that can be modelled (with different degrees of approximation) in the current state of the art. In the context of this equifinality oriented study, I think that potentially important processes that are not or incompletely modelled should be introduced.
AC: Done.

3 - p17776 l10-11: The Buizert et al. (2012) study, which details a method of data uncertainty evaluation, states at the bottom of page 4268 (first column) that these uncertainties should not be interpreted in an absolute sense. I think that data uncertainty is not so easily dealt with and deserved more comments. For example, possible calibration biases between firn data and atmospheric scenarios are not easily detected for species with strong concentration gradients in the upper firn. This argument may somewhat oppose to the principle of using mostly/only species with strong concentrate on gradients in firn to constrain the diffusivity.
AC: We agree that data uncertainties are not easy to estimate or deal with. We have added more discussion of data errors.
Calibration biases, which are often systematic rather than random, are certainly difficult to deal with. We introduced our Synthetic B case (that has systematic errors that are comparable to calibration biases seen in real observations) to demonstrate the effect of these errors

compared to the idealised Synthetic A case. In the end, the question of whether possible calibration biases will outweigh the benefits of tracers with strong concentration gradients in the firn will be a question of the relative size of the signal versus the noise (including calibration biases), in determining which tracers are more useful, in addition to using as many tracers as possible in the hope of reducing the impact of any calibration bias.

4 - p17777 l1-5: A reference provided by the authors (Beven, 2006) presents the estimation of model errors as essential in the equifinality technique. This issue is a major concern for inverse modelling techniques in general and several methods have been developed to estimate such errors (some are discussed by Beven, 2006) and should be introduced here.

Model error is never easy to deal with in inverse modelling. We do already discuss model error in the introduction and take it into account in our choice of the range of solutions to accept, and also by using alternative model formulations for processes (convective mixing, dispersion). We don't believe that the paper would benefit from introducing the methods discussed by Beven 2006.

5 - p17779 l25: If I understand well, there is at least one exception to the monotonically decreasing diffusivity with depth: the melt-layer case at DE08. Is the monotonically decreasing assumption necessary with the genetic algorithm? If no, what is its advantage?

AC: The monotonic assumption is not necessary for the genetic algorithm, but is used here for the following reason. As Rommelaere et al (1997) pointed out, many diffusivity profiles are able to reproduce the firn data, but most of them are unrealistic and characterised by large oscillations. Some kind of smoothness constraint is therefore needed for an automatic method of diffusivity estimation. We use monotonicity as our smoothness constraint, where the monotonic assumption ensures that we avoid diffusivity solutions with oscillations. Melt layers are treated as a special case, as described in the Supplement and also now mentioned briefly here. See also the response to the first comment by Reviewer 2 below, for physical reasons for the monotonic assumption. These reasons are now described more clearly in the paper.

6 - p17780 l10-13: I am puzzled by the irregular shape of the upper bound diffusivity uncertainty range on Figures 8, 12 and 13. Could it be related to the prior selection of solutions?

AC: It is related to the choice of points for interpolation with the monotonic spline (fixed diffusivity values for which the corresponding open porosity is estimated, see Fig. 1), and in particular the point at $400 m^2 yr^{-1}$. Increasing the number of points would lead to smoother upper bounds, but rapidly increase the computational burden for the genetic algorithm. Our choice of 16 points was a balance between having enough points to define a range of accepted solutions, and computational constraints.

7 - p17780 l21 - p17781 l3: How is the melt layer case handled together with the monotonically decreasing diffusivity assumption?

AC: It is handled as a special case, with diffusivity reduced by a fixed amount across the boundary between two particular model layers that corresponds to the depth of the melt layer as it moves with the ice. This was described in the Supplement section 1.9. It is now mentioned in the main text with the discussion on the monotonic assumption.

8 - p17781 l11-13: How genetic concepts such as "breeding", "mutating" and "generation" apply to firn should be introduced.
AC: They are genetic algorithm concepts. The parameter sets (in our case, the parameters are open porosity values corresponding to the fixed diffusivity values, as well as surface diffusivity, well-mixed layer depth, eddy diffusion parameters etc) are combined (bred) and altered (mutated) to create a new generation of parameter sets from the previous one, analogous to genetics. This is now explained more clearly in the text.

9 - p17781 l25- 17782 l4: The error range chosen for the definition of equifinality seems somewhat arbitrary to me:
- Although it is very close to the range covered by the 6 models in Buizert et al. (2012) (0.73-0.92 for NEEM-EU), taking into account the new CSIRO model performance, 3 models show results within a very narrow range (0.73-0.74). Buizert et al. underline the fact that the OSU model has limited degrees of liberty to tune the diffusivity, thus the 0.73-0.80 range could be more significant than the 0.73-0.92 range.
- A more pessimistic view of the Buizert et al., 2012 results could come from the fact that no model fits all data points within error bars. A spectacular example is the HFC-134a data point at _65m depth, which is not matched by any of the models. Therefore I think that the current state of model and data error understanding in the firn modelling community does not allow to clearly define a single robust criterion for equifinality. Could a narrower range of solutions (e.g. 0.73-0.80) be easily re-selected and commented (e.g. compared with Fig.8)?
AC: We agree that there is no single definition for a suitable threshold, and therefore no correct choice; we are guided by the F-ratio test relating thresholds to confidence levels (see next comment) as well as to some extent the spread of results from the Buizert et al study for NEEM. From the F-ratio test, the threshold of 0.92 at NEEM corresponds to a confidence level of almost 100% (for our choice of 16 parameters), while 0.80 corresponds to a confidence level of about 98%. When now use a 68% confidence level for all calculations (for NEEM this corresponds to a threshold of 0.78), and we also show results for some alternative thresholds in a number of cases, showing that our conclusions are not strongly dependent on the choice of threshold. We agree that the Buizert et al study and the new CSIRO model would suggest that firn models are now getting their best estimates around Phi=0.80 or lower, if we exclude the OSU model for reasons given in Buizert et al. However, as we said at line 4 on page 17782, care is needed to avoid retaining only the very best solutions, in order to allow for uncertainty due to model error, unmodelled processes etc. With any of these choices of threshold for NEEM we are fitting to the full set of observations better than expected by the data uncertainties (i.e. Phi is less than 1.0).

10 - p17782 l8 - p17783 l8: the method described is not used, thus this one page long discussion could be removed. p17783 l1: why 16 parameters?
AC: The F-ratio test relates thresholds for Phi to confidence levels, and is used for all our calculations to guide our choice of threshold for Phi. We retain the discussion.
The 16 parameters are the 14 diffusivity levels given at on page 17780 at line 11, plus surface diffusivity and well-mixed layer (convective zone) depth. This is now clarified.

11 - p17784 l9-11: Why is the LGGE-GIPSA model diffusivity used to generate the "true concentrations" of a different model? As noted by e.g. Buizert et al. (2012), at the beginning of their Section 4.1, the spread of diffusivity profiles from different models reflects in part differences in model physics, which are compensated for by the

diffusivity adjustments. I find the use of the word "true" in the synthetic scenarios A and B context confusing.
AC: We had to use a diffusivity profile to generate observations for our synthetic cases, and instead of inventing something ourselves, we chose to use a diffusivity profile that would give us NEEM-like observations. Although differences in model physics mean that a diffusivity profile used in two different firn models will give different results, as pointed out by the reviewer, we wouldn't expect the results to be vastly different. Furthermore, the synthetic data don't need to be identical to NEEM. The LGGE-GIPSA model diffusivity from the NEEM intercomparison was chosen for 2 reasons: unlike the other models in that study, LGGE-GIPSA didn't include dispersion in the lock-in zone, so we could create NEEM-like concentrations using only molecular diffusion, and it was readily available in the Supplement to Buizert et al. We are not trying to recreate the results that would be given by the LGGE-GIPSA model, except in an approximate way. We are not comparing the results of the synthetic cases with the actual NEEM observations. This has been explained more clearly in the paper.
By "true" diffusivity, we mean the diffusivity from the LGGE-GIPSA model that we use in the CSIRO model to generate what we call the "true" concentrations. We then try to recover the "true" (i.e. the LGGE-GIPSA) diffusivity. This is now clearer.

12 - p17784 l12-15: I'm not sure to understand how noise is added to the "true concentrations": is it just used to define sigma_i or are the "true concentrations" really shifted in a trace gas specific way? If yes, is this shift chosen as a single selection of the random variable or are several selections used? I do not see any obvious noise on the "observations" on Figures 2, 3 and 6. It would be interesting to plot the sigma_i values used and comment how they compare with the diffusivity-driven envelopes. What motivated the choice of a noise scaled to the RANGE of each tracer? I do no see a connection between this assumption and some characteristic behaviour of experimental/model uncertainties.
AC: Before it is added to the "true" concentrations, the random noise is shifted and scaled so that the noise has zero mean and sd equal to 0.5% of the range of each tracer. Due to the small number of pseudo-observations, the law of large numbers does not apply. Without this normalization step the noise we generate might not have the desired mean and SD. We have clarified this in the text.
We have added a plot that compares the sigma_i, model-data mismatch and diffusivity-driven envelopes for the NEEM real data case, as suggested at comment 27 below. We do not wish to also add a similar plot for the synthetic case, but instead have indicated sigma_i in the figure for calibration with individual synthetic A tracers.
The choice of noise scaled to the range of each tracer gave us a way to compare tracers based purely on the shape of their atmospheric history relative to the sampling date, without conflating this with the different data uncertainties that each tracer has (due to how accurately they are measured, possible calibration bias, uncertainty in atmospheric history etc). Clearly both aspects are important, as we have mentioned throughout the paper, but in our synthetic tests we wished to focus on the impact of the atmospheric history, in order to demonstrate our method and look at the principles behind the choice of which tracers are best. After the synthetic cases, we go onto look at real cases with real uncertainties. We have added discussion of the motivation for our choice.

13 - p17784 l23-25: is a single random number value tested or are several values used in several simulations to make the test statistically significant?

14 - p17785 l2-3: does the choice of random values affect the estimation of which tracer or subset of tracers is best suited?
AC: For Synthetic A, we tested two simulations of noise (only one is presented here), and the results were very similar. For Synthetic B we only tested one, and we would expect the results to be dependent to some extent on the choice of random values, but we do not use them to rate individual tracers, rather we wish to see how a single case with systematic errors differs from Synthetic A. The fact that the relative constraint provided by the different subsets of tracers is similar for the Synthetic A and B cases suggests that the atmospheric history is having a stronger impact on the results than the data uncertainties for the levels of uncertainty that we have considered. While we agree that several simulations would make the results more statistically significant, this is not really what we are after.

15 - p17786 l1-5: How different is the DE08-2 data based closed porosity profile from the modified Goujon et al., 2003 parameterisation used in Buizert et al., 2012? How does it affect the results? What is used for NEEM?
AC: The closed porosity depth profiles from the parameterisation (and parameters) used for NEEM by Buizert et al, and the DE08-2 data based closed porosity profile are quite different. The increase in closed porosity (decrease in open porosity) occurs for lower values of density at DE08-2 than NEEM, but the density value corresponding to zero open porosity is fairly similar. It is difficult to match both the increase in closed porosity and the value of zero open porosity (where air can no longer be extracted) in the DE08-2 closed porosity data simply by tuning the close-off porosity parameter in the Goujon et al., parameterisation.
The choice of closed porosity parameterisation is not expected to affect the modelled diffusion processes in the open pores significantly. We are tuning the diffusivity, so any difference in open porosity will be compensated for by a difference in diffusivity. However, it is likely to affect the bubble trapping (which is not the focus of this study). In tests now briefly described in the paper, we calibrated DE08-2 diffusivity using 3 different spline fits to the DE08-2 closed porosity measurements (each used different corrections for cut bubbles), and the Goujon et al parameterisation with two different close-off porosity parameters (one to match the increase in closed porosity but giving too shallow total close-off at 83m, the other giving too deep an increase in closed porosity but giving a better total close-off depth of around 90m). This gives us a way to include the uncertainty due to closed porosity in any modelled quantity. However, assuming the closed porosity measurements are reliable, this range of closed porosity options most likely overestimates the uncertainty due to closed porosity, particularly as the Goujon parameterisations do not match the closed porosity observations well.
For NEEM we used exactly the parameterisation used in Buizert et al., 2012, as described in Section 2.3.1 (we use the physical firn characteristics from Buizert et al).

16 - p17787 l6-13:
- l6-12: I guess that Law Dome ice core data are used to constrain the early part of the atmospheric scenarios for other species such as CO2 and CH4, the atmospheric scenario estimate and firn diffusivity evaluation are thus somewhat inter-dependent. But if such tracers are also excluded, there will not be much remaining data to constrain the deep firn. Could you comment?
AC: Yes, Law Dome ice core data are used to create the CO2 and CH4 atmospheric histories that drive the firn model. However, Law Dome ice core CH4 and CO2 are dated using a constant offset between the age of the air and the age of the ice (the age of the ice is quite accurately known), and the ice core record overlaps the modern atmospheric record at the

recent end and consists of data from 3 different cores. Therefore, there is not a strong circularity here, and we don't believe there are grounds to exclude this data.

- l12-13: This would be an interesting test to perform on NEEM.
AC: We now comment on a case for NEEM that excludes the firn data corresponding to emissions-based atmospheric histories.

17 - p17788 l27 - p17789 l1 : I do not understand this sentence - gravitational separation is driven by partial pressure gradients (see e.g. Schwander, 1989), why is it affected by the use of Deddy rather than Dmolecular? Trudinger et al., 1997 do not use dispersive mixing and seem to correctly simulate d15N2 in the lock-in zone (Figures 4-6). Schwander et al., 1993 also simulate a d15N2 flattening in the lock-in zone without Deddy (Fig 4).
AC: Deddy affects all gases similarly, and therefore counteracts the molecular separation due to gravitational fractionation, this is not the case for Dmolecular.
In the examples cited by the reviewer, the absence of gravitational separation in model results is due to using small or zero molecular diffusion in the lock-in zone. We found in our calculations for NEEM (see Supp line 572) that with Dmolecular (and no Deddy) in the lock-in zone, d15N2 increases at a very slow rate but is just as consistent with observations as constant levels that would be obtained with Deddy in the lock-in zone. Our main aim here is, given that some scientists believe dispersion may occur in the lock-in zone, to include this uncertain process in our estimate of equifinality. In most cases we can fit the observations equally well with and without it, and therefore we cannot make any statements about how likely it is to be occurring. DSW20K is the exception, where we find improved fit to observations with eddy diffusion in the lock-in zone.

18 - Section 3.1.1 : some choices of vocabulary render this section difficult to understand in detail at first reading. I had to take some notes on definitions given in this section, section 2.3.2, Table 2, Table 3 and Figure 2 caption to understand the main results.
AC: We have now defined these terms (true, observations etc) at the start of the section, and tried to be more consistent with their use to avoid confusion.

- The word "observations" is used for synthetic data with or without added noise
AC: Observations should refer to synthetic data with noise – this has been corrected.

- The definition of the "true" solution is confusing. For trace gas mixing ratios, is it the LGGE-GIPSA model results or the CSIRO model results with the LGGE-GIPSA model diffusivity ? How different are these two solutions?
AC: The "true" solution refers to the CSIRO model results with the LGGE-GIPSA model diffusivity from Buizert et al. This is now explained more clearly. We have not compared these two solutions, they are broadly similar (compare Figs. 2 and 8) and small differences are not important for our study (see also response to comment 11 above).

- I guess that "true concentrations" at p17790 l8 refer to synthetic data without noise rather than experimental data, I'm I right?
AC: yes

- The fairly systematic use of subset numbers in the discussion of which group of tracers is best suited to fit itself and other tracers is confusing for someone who does

not yet remember which species is included in which subset.
AC: We have tried to mention which tracers are added from one subset to the next to improve readability.

I also had to read several times Sections 3.1.1, looking at Table 3 in parallel and taking some notes to try to understand what the main conclusions are. In my view, they are:
AC: We have tried to make the main conclusions clearer.

- diffusivity fairly well constrained with CH4, d15N2 and SF6. As d15N2 shows little sensitivity to diffusivity (p17790 l11-13), would CH4 and SF6 be enough?
AC: see point 19 below

- 5 tracers are needed to avoid over-fitting the data (which is quite a lot if the fact that those tracers have reduced uncertainties compared to experimental data is considered) I think that the main advantage of the methodology for building synthetic scenario A is that it allows to properly detect an over-fitting of the data when not enough tracers are used. In my opinion, the discussion of results should be shortened and focused on that point rather than on the ranking of tracers which is directly affected by the wrong amplitude of the uncertainties.
AC: As discussed in response to other comments, we are interested in the value of different tracers due to their different atmospheric histories, without the complications of real errors, this was the basis for these calculations.

19 - On p17784 l11, it is said that synthetic A scenario uses a 3.66m well mixed layer, and on p17790 l11, it is said that d15N2 is quite sensitive to the depth of the well mixed layer, could you clarify?
AC: The forward run of the firn model to generate the "true" concentrations used a 3.66m well-mixed layer. Modelled d15N2 varies as different values of the well-mixed layer depth are tested by the genetic algorithm, but hardly changes as different molecular diffusivity profiles (above the lock-in zone) are tested. The consequence of this is that d15N2 is a useful constraint for the well-mixed layer depth, but quite useless for constraining the molecular diffusivity above the lock-in zone. We have clarified this point in the paper. As asked above (would CH4 and SF6 be enough?), d15N is useful in addition to CH4 and SF6 to help constrain the well-mixed layer (and also probably helps constrain the lock-in depth).

20 - p17791 l26 - p17792 l2: how is the "true" depth of the well mixed layer determined? What is the interplay between this depth and diffusivity? In my view, there is a conflict between the mixed layer concept and the correct simulation of species which show concentration gradients in the near surface firn (CO2, SF6, HFC-134a) that is in part hidden by the fact that synthetic A scenario does not use the surface "data" points and the depth resolution used. I doubt that d15N2 data not corrected from the effect of thermal diffusion would be consistent with a well-mixed layer.
AC: We chose the "true" depth of the well-mixed layer to be 3.66m, this gives variation in concentration near the surface that is roughly consistent with NEEM observations.
Tracers in the CSIRO model are completely well-mixed down to the depth of the well-mixed layer, then transported by molecular diffusion below this depth. As molecular diffusivity is not used through the well-mixed layer, it has no affect on modelled concentrations – a consequence of this is that molecular diffusivity is not constrained over the range of the well-mixed layer.

These comments relate to the section on synthetic experiments. Synthetic A does not use the surface data points (i.e. 0m depth) in the calibration, because the surface concentration cannot vary with the diffusion parameters, it depends only on the value of the atmospheric history at the sampling time, so is of no use in our calibration. Below this we use the same measurement depths as at the NEEM 2008 site. The synthetic calculations do not include the effect of thermal diffusion, so this is not relevant to these synthetic calculations.
In a real case, surface data points will also not be affected by diffusion parameters in the firn, so are not directly useful to an automatic calibration, however, any difference in the surface firn concentration from the atmospheric history at sampling time could give insight into possible errors in atmospheric history (e.g. calibration scale, spatial gradients in the atmosphere etc), and is also useful to compare with subsurface samples to determine concentration gradients. Not all sites have significant thermal diffusion.
We found that the well-mixed layer gave reasonable agreement with measured tracers at DSSW20K (doesn't appear to have been impacted by thermal diffusion) and NEEM (data corrected for thermal diffusion), but slightly better results were obtained using exponentially-decreasing "eddy" diffusion.

21 - Section 3.1.2:
- I have again a difficulty to understand the main conclusions of this section. For example the worst tracer appears to be d15N2 on p17792 l8, 14CO2 on p17793 l1 whereas p17793 l5 designates 14CO2 as the best constraint in the lock in zone. The "best tracers" for the upper firn: CH3CCl3 and HFC-134a (?) were emitted in the atmosphere only recently and have concentrations close to zero near and below the 14CO2 peak, which renders them nearly useless in the firn depth range where 14CO2 is useful. They are thus quite obviously more complementary than better or worse.
AC: Yes, we see them as complementary rather than better or worse. The discussion has been reformulated to reflect this point better, and to make the main conclusions clearer.

- Witrant et al. (2011) performed the same exercise using experimental data and uncertainties rather than synthetic data. Does this change the main conclusions?
AC: Witrant et al only show their best solution for each case, not ranges as we do, and we haven't done the individual tracer experiments with real data, so we can only compare our best cases using synthetic data to Witrant et al.'s using real data. When we do this, some of the conclusions are the same, others are different. We have added a paragraph describing this comparison to the paper.

- On Fig. 4, results with CH3CCl3-only look better than (or in a few cases similar to) results with 10 tracers in the 15-60m depth range except for d15N2. Could you comment?
AC: We now plot results for a lower threshold, and the case with 10 tracers is mostly better than or sometimes as good as the CH3CCl3 case, but with the 1.125 threshold the CH3CCl3-only case did often have smaller spread than the case with 10 tracers. CH3CCl3 is a very good constraint on diffusivity in the 15-60m range, and lacks information on the well-mixed layer and lock-in zone that is provided by other tracers like d15N and 14CO2. To understand why the CH3CCl3-only case had lower spread than the case with 10 tracers over the 15-60m range, it is important to think about how the fit to calibration observations can vary below the Phi threshold. With 10 tracers, it would be possible to fit observations of some tracers in the lock-in zone very well to offset a slightly greater mismatch with other tracers through the 15-60m depth range. The case with only CH3CCl3 cannot do this.

- On Fig. 4, d15N2-only constrained diffusivity leads to very poor results for all tracers except d15N2 itself, which is also poorly constrained by all other tracers compared to itself. Does it mean that there is a conflict (data inconsistency or model's wrong representation) between d15N2 and other datasets?
AC: No. Because this case uses synthetic data we can rule out both a data inconsistency and model's wrong representation. We could not make this conclusion if we had used only real data. d15N2 and the other tracers contain complementary information, with d15N telling us about convective mixing near the surface and possibly dispersion in the lock-in zone, and the other tracers telling us about molecular diffusion. This is the reason for including d15N in our choice of the three best tracers.

- I would prefer to see the equivalent of Fig 4 results/analysis for the simulations constrained with "real" experimental data than scenario A synthetic data.
AC: We haven't run the single tracer cases with real experimental data and uncertainties – there is only so much we can do, and our choice is to do this with the synthetic data, for the following reasons. It allows us to focus on the value of different tracers due to their atmospheric histories alone. The synthetic results from the different subsets of tracers are interesting enough just due to the atmospheric histories, and adding the effect of the errors and inconsistencies in real data may obscure some of the key messages, like the value of CH3CCl3 if we can reduce the errors and inconsistencies, and the complementary nature of some tracers.
Each site and firn sampling campaign will differ in the degree to which errors and inconsistencies affect the results. For example, offsets between firn measurements and atmospheric histories are likely to be lower in the southern hemisphere than the northern hemisphere because of the smaller spatial gradients in many long-lived tracers; offsets will also be lower if the firn measurements are made in the same laboratory as the atmospheric history measurements, with further benefit if the atmospheric history is based on an atmospheric air archive (such as the Cape Grim Air Archive) measured all at one time. In addition, different firn column depths and sampling dates will alter which tracers have the strongest gradients in the firn. Therefore, we see value in looking at the principles behind what makes a good choice of calibration tracer, on which to build when we add in the effect of real data errors that will differ to some extent from one study to the next.
Clearly there would also be value in doing the single tracers experiments with real data, but our choice here has been to use the synthetic data, and we have demonstrated the benefits.

22 - Section 3.1.3:
- The left side of Fig.5 is not discussed and could be removed
- I do not understand why the subset d15N2 + CH3CCl3 + 14CO2 is discussed only in terms of spectral widths and not in terms of PHI_A etc. (it was not presented in Section 3.1.1)
- Besides the above new subset, I do not see what new conclusion is brought by the spectral width analysis. Could this Section be suppressed, the d15N2 + CH3CCl3 +14CO2 subset presented in Section 3.1.1, and the spectral width analysis shown only for "real" data (Fig 14)?
- I do not understand why much more detailed results are shown and presented for scenario A than for the simulations using "real" experimental data
AC: We have removed all but panel (o) of Fig 5, and combined this with similar panels in Fig 7. The d15N2 + CH3CCl3 +14CO2 subset was determined with the benefit of hindsight as the minimal best combination, based mainly on the tests of individual tracers and the initial

combinations of tracers tested, so was not included in the original subsets. We now introduce it following the individual tracer runs, rather than in the discussion on spectral width.
The spectral width analysis for Synthetic A has been shortened, to have a similar amount of detail as for the other datasets. We do present a comparison of the range in spectral width for Synthetic A for two choices of PhiA, which is a useful sensitivity test, and helps address the reviewer's comment 9 above. We do not wish to completely remove the spectral width analysis from the synthetic section.

23 - p17794 l24-25: the "similarity" between synthetic B and "real" data has already be commented in Section 2.3.2. However it misses an element of primary importance: in error (1) sd can be very different from one tracer to another. For example the three species that have non null concentrations in deep firn (CO2, CH4 and 14CO2) have very different uncertainties and this is likely the primary driver for their ranking as efficient tracers.
AC: The characteristics of the errors are similar to the real data, not the specific values. We have now made this point in the paper (in what was Section 2.3.2). The similarity between the results obtained with the Synthetic A and B and real NEEM data suggests that the atmospheric history is still an important determinant of the value of different tracers. Clearly data error varies from one tracer to another, and if you wanted to know which tracers were most valuable for a future firn campaign you would need to take them into account. Our main reason for considering synthetic B is to take the next step from synthetic A by adding larger, correlated errors, while retaining the benefits of synthetic data.

24 - Section 3.2:
- The reason why values in column "truth" of Table 3 for synthetic scenario B deviates from 1 is not explained. I guess that the main reason is that a single value of the "random" calibration bias was used for each tracer, I'm I right? To which extent deviations from 1 in column "truth" reflect a lack of statistical significance of the results?
- The analysis of synthetic scenario B results seems rather inconclusive to me. An important point to me in this respect is the fact that 5 tracers were already needed to avoid an over-fitting of the selected tracers in scenario A and there is no subset having more than 5 and less than 10 tracers in scenario B.
- Some differences between scenario A and scenario B have the potential for helping to answer important questions: about the effects of depth resolution of the data, and about firn data versus atmospheric scenario calibration bias but are not tested individually. Thus I think that scenario B related description and analysis could be suppressed without affecting the main conclusions of the manuscript.
AC: Yes, the deviation from 1.0 depends on the single realisation of systematic error.
The main results from the Synthetic B cases are a) the increased spread of solutions when larger, correlated errors are used, compared to Synthetic A, and b) the similarity of synthetic A, B and NEEM calculations for the different subsets of tracers. If the Synthetic B calculation had not been included at all, there could have been the criticism of Synthetic A because it included only small, Gaussian errors, and was therefore unrealistic. Synthetic B is an intermediate step between Synthetic A and real data. Computational constraints prevented us from exploring many realisations of synthetic data error, however, the single Synthetic B realisation is still useful relative to the Synthetic A case. We didn't look at the effect of depth resolution, or calibration bias, as there is a limit to how many cases we can cover.
We have significantly shortened the discussion of Synthetic B, but wish to leave it in the paper.

25 - p17796 l17 - l23: I would be very interested to read a more detailed analysis of the comparison of the new CSIRO model results with the other models in Buizert et al. (2012). For example, as the diffusivity and 10 tracers concentration results of the six models are provided in the supplementary files of Buizert et al. (2012), it is possible to build the equivalent of column "TenEddy" in Table 4 for the five other models. An interesting aspect of the Buizert et al. (2012) study is to provide a benchmark of model results to which other firn models or new versions of the same models can be compared. In this respect, it would be interesting to provide equivalent result files to those in the Buizert et al. supplement with the new CSIRO model.
AC: We have included in the Supplementary Material the equivalent results files to those in the Buizert et al. supplement for the new CSIRO model, and figure showing the results for the diagnostic scenarios from Buizert et al.
While it would be possible for us to create a table of Phi for each tracer using the results from all models in Buizert et al. as well as our new model, but we do not wish to do this here. By only comparing a single case from each model, it may be difficult to interpret the results in terms of differences between models (see next comment). With our results available in the supplement, it is now possible for anyone to do this should they wish, but we would argue that the value would be limited due to equifinality.

26 - Table 4: when comparing columns "Ten" and "TenEddy", I do not understand why 14CO2 is strongly affected by the convective zone representation in the model (it has a nearly flat profile and a low depth resolution in the upper firn). This is not commented in Section 3.3.
AC: There is essentially no difference between “Ten” and “TenEddy” in the fit to 14CO2 above 60m, all the difference occurs in the lock-in zone (where eddy diffusion from the convective zone representation is negligible). The difference in Phi for 14CO2 between Ten and TenEddy is not particularly large, and is within range of variation of Phi for individual tracers for solutions with similar overall Phi. This is an example of equifinality, and shows that a comparison of just the best solutions can be confusing. Comparison of ranges of solutions would be more useful, as is one of the main points of our study. We have added this point to the paper.

27 - Figure 8: I think that the widths of the diffusivity related envelope and the experimental uncertainties should be compared. The much larger width of the diffusivity related envelope in the 20-60m range for CH3CCl3 and HFC-134a is striking. In other cases the diffusivity related envelope is much smaller than data uncertainty (e.g. 14CO2 in the 0-60m range) - over-fitted data? This comparison is not easy to view when the envelope and uncertainties are small. Some data points remain well outside the diffusivity related envelope - outliers? How dependent are the results on the choice of the equifinality criterion ? Are the tracer/depth ranges where diffusivity related envelopes are large the same as those where the results of different models diverge (Buizert et al., 2012)?
AC: We have added a figure comparing the diffusivity related envelope (with different choices of equifinality criterion Phi), the experimental uncertainties and the model-data mismatch for our best case to the Supplement.
The tracer/depth ranges where our NEEM TenEddy case has largest envelopes (CH3CCl3 and HFC-134a between 20-60m) is not the same as where the Buizert et al models diverge – the models show larger spread in the peaks of 14CO2 and CH3CCl3 than our spread, and

more similar spread among the other tracers without CH3CCl3 and HFC-134a being clearly more spread.

28 - p17797 l1-l22: I could not find a strong conclusion in this scenario B versus real data discussion.
AC: Most of this discussion has been removed.

29 - p17798 l8-l12: The interest of HCFC-142b and HFC-43-10mee as additional tracers would be enhanced if the enlarged envelopes for CH3CCl3 and HFC-134a was due to an inconsistency between the CH3CCl3 and HFC-134a datasets (e.g. calibration bias). The consistency of CH3CCl3 and HFC-134a datasets could be tested with the "19 representative solutions": is a solution fitting well one dataset bad at fitting the other?
AC: There is a tendency for solutions with a better fit to CH3CCl3 to also fit HFC-134a well, and vice versa, but there is also a lot of scatter. Therefore there does not appear to be an inconsistency between these two datasets.

30 - Sections 3.4 and 3.5 DE08-2 and DSSW20K have a much lower sampling resolution than NEEM. An interesting test could be to reduce the resolution of the NEEM data and evaluate how the diffusivity and tracer concentration profiles are affected.
AC: We have added a case with reduced resolution of the NEEM data.

31 - Figure 10 shows model results in firn down to 100 meters depth, whereas on p17786 (l4) it is said that open porosity is zero around 90m. Could you comment?
AC: For computational convenience, the CSIRO firn model has a numerical value for the mole fraction of a tracer in both the open and closed porosity at all depths throughout the firn and ice. After all bubbles in a layer have been trapped, although the mole fraction in the open porosity corresponds to zero porosity, it continues to be advected with the ice, however it is of academic interest only. We agree that plotting firn concentrations below the depth of zero porosity has no benefit and is confusing, so have truncated the plots at this depth.

32 - p17799 l11-l14 and Figure 11:
- I do not see a clear evidence of a convective zone in the d15N2 data and model results. Could you comment? Is the d15N2 slope in the diffusive zone deviating from the barometric slope? Is the enlarged envelope on the results in the 30-40 meters range affected by the convective/dispersive mixing effect going throughout it? Section 2.5 mentions simulations with eddy-diffusion and a well mixed layer, how do they compare?
AC: With only a small number of d15N2 measurements through the diffusive part of the firn at DSSW20K, it is not easy to define the convective zone. However, we get similar estimates of its thickness from different methods: 1) the barometric equation shifted by 4.7m gives the best fit to the d15N2 measurements, 2) the model tuned with a well-mixed layer prefers a well-mixed layer depth around 4.5m, 3) with exponentially-decreasing eddy diffusion used in the model for convective mixing, we get equal eddy and molecular diffusion (used in Kawamura et al 2006 to estimate the convective zone thickness) at around 5m. The barometric slope plotted from the origin is clearly higher than all of the d15N2 observations. We take this as fairly clear evidence of a convective zone.
The enlarged envelope around 30-40m is not due to the convective/dispersive mixing going through it, as it also occurs with a well–mixed layer of around 5m and no eddy diffusion. As we pointed out in Section 3.5, the molecular diffusivity decreases by at least a factor of 10

over this depth range, and this leads to the greater uncertainty. Increased depth resolution of observations would most likely reduce the uncertainty here.

- I am puzzled by the physical meaning of an eddy diffusion term which goes throughout the firn. In the upper firn, eddy diffusion is a simplified way of representing fast mixing due e.g. atmospheric pressure variations or wind pumping. DSSW20K and DE08-2 have nearly the same temperature and pressure, thus nearly the same CO2 diffusion coefficient in free air. Comparing Figures 10 and 11, CO2 diffusivity in the upper firn is much smaller (twice to a third) at DSSW20K than at DE08-2. What is the physical meaning of a high Deddy contribution combined with a low Deddy+Dmolecular diffusivity? In the deep firn, Deddy is assumed to represent a form of transport in firn layers mostly isolated from the atmosphere (Supplement, Section 5). What is the physical meaning of a Deddy term in the diffusive zone (where gravitational fractionation occurs for d15N2)?
AC: We found with using exponentially-decreasing eddy diffusion for convective mixing near the surface at DSSW20K that the eddy diffusion is still around 0.3 m2/yr when the molecular diffusion goes to zero. This is of the order of the dispersive mixing. This now appears to have occurred because some eddy diffusion in the lock-in zone improves the fit to observations, although we note that the exponential form of the eddy diffusion is probably too simple. We have noted this in our discussion, but now use as our preferred case for DSSW20K a case that truncates the Deddy by 30m, to separate the effect of eddy diffusion near the surface and in the lock-in zone. The Deddy term in the diffusive zone is overwhelmed by the molecular diffusion flux, and therefore of little importance. DSSW20K still has a smaller Dmolecular than DE08-2 when the well-mixed layer is used, so this appears to be a difference between the sites, but we are not sure of the reasons.

33 - p17799 l15-16: deep firn CO2 is also under-estimated at DE08-2 (and NEEM) but not South Pole, could you comment?
AC: The underestimation of deep firn CO2 at some, but not all, sites is certainly a puzzle. It will no doubt be the focus of other studies in the future, and is beyond the scope of this work. We now mention this in the section on Comparison of sites.

34 - Sections 3.4 and 3.6: Witrant et al., 2011 also modelled DE08-2 and South Pole. Are there significant differences between the results of the two models?
AC: There are remarkable similarities in the structure of the diffusivity in our best case for DE08-2 and Witrant et al's diffusivity, particularly the step-like variation with plateaus at around 50m and 73m. This is interesting, given the different models, functions for diffusivity, (including the fact that we use a melt layer and they don't) and calibration observations (in addition to CO2 and CH4 used in both studies, we used 14CO2 and d15N and they used CFC-11 and CFC12) used in the two studies. This is now mentioned in the paper.
Our best cases for South Pole 1995 and 2001 also have diffusivity generally similar in magnitude and form to that estimated by Witrant et al.

35 - Fig 14:
- The uncertainty envelopes plotted refer to the 19 representative solutions rather than all solutions corresponding to confidence intervals of 68 %, why?
AC: During a genetic algorithm run, we typical find several thousand solutions with Phi below our chosen threshold. We select 20 of these to represent the full set – the best case plus 19 others with different values of Phi up to our threshold and different diffusivity in different parts of the profile, so often solutions at the edge of the full envelope are chosen in our

representative set. We saved the depth profiles of the ice properties and reference tracer concentrations for our full set of solutions from the genetic algorithm run, so we can show these quantities for the full range of solutions. However, when the model is used to calculate other quantities, such as the age distributions in Fig. 14 or the additional tracers in Fig. 9, we use only the representative set of solutions.

- The plotted envelopes are fairly symmetric at NEEM but very asymmetric at other sites, could you comment?
AC: There is no particular reason why the best solution should be in the middle of the envelope. For this reason, it is valuable to use an ensemble of solutions to represent model prediction uncertainty, rather than assuming errors are normally distributed around the best solution.

- Spectral width in deep firn seems to increase regularly with depth at NEEM and DE08-2 but an important slope break (flattening) is visible at DSSW20K, South Pole 1995 and 2001. Is it due to a different model configuration? Does it affect all solutions?
AC: We're not sure why this occurs. It seems to affect all solutions.

- NEEM and DSSW20K, which have somewhat similar accumulation rates, show similar behaviours approximately down to the lock-in depth (where d15N2 stops to increase): a slow increase of the spectral width up to 4 years, then the spectral width increases faster (this is a common behaviour of all sites, which does not surprise me), then the increase in spectral width with depth suddenly stops at DSSW20K whereas it continues at NEEM. Is this striking difference well constrained by the data or due to a difference in model configuration (e.g. eddy diffusion, sampling resolution, etc.)?
AC: The lower edge of the NEEM ranges of solutions corresponding to the 68% is roughly the same as the upper edge of the DSSW20K range, so they are not vastly different. We are not sure of the main reason.

36 - p17801 l5-l11: A simple test could help being more precise in the comparison between the two South Pole drill sites: using the diffusivities calculated for one site to simulate the other site.
AC: This would need to be done for the ensemble of representative solutions in order to be a useful comparison. However, we have decided not to do this comparison, and prefer instead our comparison of the ranges of spectral width for the CO2 age distribution.

37 - Section 3.7 I am very surprised that the title question : "How well do different tracers constrain the firn diffusivity profile?" is not discussed in this short "Comparison of sites" section. For example, the smallest set of tracers available is CO2, CH4, SF6 and d15N2 (for South Pole 2001). These four tracers are available at all other sites. How degraded are the results at other sites if only those four tracers (with "real" data and uncertainties) are used to constrain diffusivity?
AC: There are so many things we could have tested with this method and the available data, and while this suggestion would be an interesting test, the paper and Supplementary Material are already long, so we have not included it here.

38 - Section 3.8: The question raised in this Section: how are the results affected by a possible bias in diffusion coefficients? is very interesting but I have difficulties to understand the meaning of results and their discussion in detail. The main conclusion(s) of this section is(are) not clear to me.

AC: This section has been moved to the Supplementary material to simplify the paper, and we have tried to present the main conclusions more clearly.

***Synthetic A tests:**
- p17802 l3-l8: As the synthetic data were generated using "true" diffusion coefficients and the firn model, allowing the diffusion coefficients to deviate from the "true" value can only deteriorate the results, right?
AC: Yes.

- Table 3: comparison of columns PHI_A (or PHI_At) Ten and TenDC for CFC-11 and CFC-12. I do not understand why CFC-11 seems quite affected by variations in the diffusion coefficients whereas CFC-12 is much less affected. These species have very similar atmospheric histories and fairly similar diffusion coefficients, CFC-12 concentrations are about twice higher than CFC-11 concentrations thus the synthetic A uncertainties proportional to the species range of concentrations affect more CFC-12. Is this due to a lack of statistical significance of the results (p17802 l6-l7: "it has become harder for the GA to locate the true solution")?
- Table 3: for HFC-134a, why is the difference between Ten and TenDC very large for PHI_A but very small in terms of PHI_At?
AC: With uncertainties proportional to the range for each tracer we expect that, on average, CFC-11 and CFC-12 would be affected similarly by their uncertainties. However, as we consider only one realisation it appears that the best case is further from the truth for CFC-11 than CFC-12. Phi for different tracers has a range that is not captured by the best case values in the tables, meaning that differences of 0.1-0.2 may be due to the range rather than real differences between cases. We have now removed the columns relating to the DC cases from the tables, as detailed comparison such as required for these questions should consider the range. Instead we discuss the results quantitatively.
The point about it having “become more difficult to locate the true solution” refers to the larger number of parameters we are now estimating, rather than the statistical significance. This point is now made more clearly.

- The differences between "real" and optimal (model adjusted) diffusion coefficients (e.g. in %) should be provided in Tables 3 and 4.
AC: This can be seen in the figures in the Supplement. We don’t believe that it is worth adding just the best case diffusion coefficients to the tables.

- Section 3.1.1 (Synthetic A results) concludes that 5 tracers are needed to avoid overfitting the data. Optimising 7 diffusion coefficients together with the diffusivity profile may thus be an under-constrained problem. Could you comment? Has the calculation been repeated twice (with a different set of random parameters) to check the stability of the results?
AC: Yes, it is most likely an under-constrained problem. We have noted this in the revised manuscript. The calculation has been done only once.

- p17802 l15-l18: If I understand well, only the solution with the lowest PHI_A (with diffusion coefficients between 0.9% too low and 2.4% too high) is considered to conclude that diffusion coefficients are well constrained in the synthetic A case. How are the relative diffusion constrained if all "equifinal" solutions are considered?

AC: There is a clear minimum for each diffusion coefficient, and it is this that we are commenting on. The scatter plots in the Supplement show how Phi for each tracer varies with diffusion coefficient.

- Due to their peak shaped atmospheric scenarios, CH3CCl3 and 14CO2 have diffused into and back out of the firn in the past. Are their diffusion coefficients significantly better or less well constrained than those of other species?
AC: We didn't estimate the diffusion coefficient for 14CO2. CH3CCl3 doesn't seem to have been any differently constrained because of diffusing into then back out of the firn. With our choice of errors proportional to the concentration range in the synthetic cases, CH3CCl3 has more variation within this range than the other tracers, because the depth profile varies from almost the lower end of the range up to the top and back to the lower end. In our case, this is likely to have the most effect.

*** Synthetic B tests: In this more complex case, I have strong doubts about the statistical significance of the results: the way a change in diffusion coefficients can correct for the synthetic B biases on the data likely depends on both the choices of the species dependent systematic errors and the shape of the atmospheric scenario. Synthetic B results for "TenDC" are only briefly commented, thus I think they could be removed from Section 3.8 and Table 3.**
AC: We have moved all Synthetic A and B TenDC calculations to the Supplement, to improve readability of the main paper. However, we retain brief discussion of the Synthetic B TenDC case. Although we agree that the results would depend to some degree on the shape of the atmospheric scenario, we believe it is still useful to consider a single case as we have done here. CFC-11, with the largest systematic errors of the cases considered, has the worst estimate of relative diffusion coefficient, consistent with expectation. Since the synthetic data are generated using the measured diffusion coefficients we already know what the answer should be. This experiment therefore only serves as a reference case for the diffusion coefficient tests on the real data.

*** Results with "real" NEEM data. The authors say at p17802 l27-l28 that the reduction in mismatch is highest for SF6 and CFC-12. In the case of CFC-12, Figure 8 suggests to me that most of the mismatch occurs in the upper firn (above the lock-in depth) and could be due to a slight calibration bias between the atmospheric scenario and the firn data. Could you comment? What is the optimal PHI_N if the CFC-12 diffusion coefficient is not adjusted? I am surprised that the CH4 diffusion coefficient is strongly modified as the CH4 data are very well fitted, with a narrow envelope, on Figure 8. Does the change in CH4 diffusion coefficient aim at better fitting CH4 or better fitting other species together with CH4 (which species, in which depth range)?**
AC: There does appear to be a slight calibration bias with NEEM CFC-12 in the upper firn, and this has probably affected the result. We haven't tried a case without adjusting the CFC-12 diffusion coefficient, and don't believe that this would be worth spending time on. We don't know why there is such a big change in the CH4 diffusion coefficient.

*** Results with "real" DE08-2 and DSSW20K data. Compared to NEEM, DE08-2 has less available tracers: only five (just what is needed to avoid over-fitting the data with the synthetic A reduced uncertainties). DSSW20K has eleven tracers but a much lower sampling resolution than at NEEM. Could you comment on how well the solution is constrained? (e.g. based on optimal versus all equifinal solutions). I do not understand**

why these sites were preferred to South Pole 1995 for the estimation of diffusion coefficients.
AC: In all cases, there is a clear preferred value for the relative diffusion coefficient, and it is this that we have been comparing in this discussion. It is possible to see how Phi varies with the diffusion coefficients in the scatter plot figures for the synthetic cases. The real NEEM case has variation of diffusion coefficients with Phi similar to synthetic B. It is difficult to comment on how well the solution is constrained, as it is difficult to distinguish between overfitting systematic noise and correcting an incorrect diffusion coefficient. The synthetic calculations give us reason for caution.
We didn't include South Pole 1995 firn in this calculation because we exclude the lower firn measurements for a number of tracers because they are based on emission-based atmospheric histories, and this accounts for a greater fraction of the firn column at South Pole than other sites (as it was sampled nearly 20 years ago, and South Pole firn contains older firn air). We also had greater uncertainty in the calibration scales of the firn measurements than for the sites that were measured in the CSIRO laboratory.

39 - p17803 l27 - p17804 l6: I agree with the authors on the fact that the proposed test would be very valuable, especially for NEEM. I would be very interested to see how the results shown on Figure 8 would be modified, and how they compare with the range of model results from Buizert et al., 2012.
AC: We have already shown what our best case Phi would be (0.66). We haven't run this calculation to calculate the ranges, as the paper is already very long and we have to draw the line somewhere. We will probably consider it in the future.

40 - Section 3.9 and Supplement Section 5. At this stage I am not able to make an opinion on the fact that a significant conclusion could be reached on dispersion in the lock-in zone for the following reasons:
AC: We have put all dispersion results into the Supplement, and shortened the section considerably to present only the DSSW20K diffusive fractionation calculation. We hope that this makes it easier to understand our conclusions. As we stated earlier, based on our results we cannot say whether including dispersive mixing improves the results in a statistically significant sense.

- I do not understand the model formulation of this process in equation (30) of the Supplement: how dispersive mixing can change the mass of the trace gas species into the mass of air (see also comment 17)?
AC: See responses to comments 17 and 57.

- Buizert et al. (2012) conclude at the end of their Section 4.1 that models with completely different parameterisations of the lock-in zone reproduce NEEM observations equally well. The first test in Section 3.9 (p17804 l10-15) leads to the conclusion that other sites are insensitive to the addition of a Deddy term in the lock-in zone. Thus unless a significant improvement in terms of PHI_N can be reached at NEEM (with real data) by adding eddy dispersion in the lock-in zone (this test is not performed), I do not see how a clear conclusion can be reached.
AC: We have now removed this sensitivity test from the Supplement.

- p17804 l20-28: I do not understand the aim of the "synthetic A" test - is this to determine if the solution is not under-constrained in a reduced uncertainty frame? What can be concluded for the real data case?

AC: Yes, that was the aim of the test, however we have now removed it as mentioned above.

- p17805 l1-l12: Why DSSW20K is chosen for this test if the above results indicate that NEEM is the only site sensitive to Deddy? I guess that the reason why 13CO2 is not used to constrain diffusivity is that an atmospheric scenario not dependent on firn/ice data (and thus Deddy) cannot be built, then is it possible to constrain Deddy at DSSW20K?
AC: We chose DSSW20K for this test because we are interested in the diffusive fractionation at DSSW20K for our firn work (we now say this in the text). The reviewer is correct, we don't have a d13CO2 atmospheric record that is not dependent on firn and ice data. Comparison of d13CO2 or d13CH4 at a range of firn sites with different characteristics may provide insights, but uncertainties may be too high at present for this to be useful now. It is not just a question of constraining Deddy, but if it turns out that there is robust physical evidence supporting dispersion in the lock-in zone, then it cannot be ignored and should be included in uncertainty analysis.

- p17805 l13: It seems quite obvious to me that allowing for dispersion increases equifinality as it adds one parameter to adjust per model depth level. From an equifinality point of view, is there any hope that a dispersion term formulated as a depth-dependent Deddy flux can be adequately quantified?
AC: Adding depth-dependent Deddy need not add one parameter per model depth, it can be parameterised more simply than that (e.g. as a first estimate it can be held constant throughout the whole firn, or here we have used 4 parameters). What is needed is better understanding of the physical processes in the lock-in zone, firstly to confirm if dispersion exists and then if it does, to put the best uncertainty bounds we can on it, to limit the equifinality. Our aim here has been to demonstrate the effect of the current uncertainty in understanding of lock-in zone diffusion, and the effect this has on diffusive isotopic fractionation.

-p 17805 l14: what clear evidence that dispersion really occurs in firn do we have? From the introduction of supplementary Section 5 and the reference cited, my impression is that:
+ Severinghaus et al. (2010) included eddy-diffusive fluxes in their model as an alternative to no diffusion at all in the lock-in zone, but what all models seem to really need in Buizert et al. (2012) is MOLECULAR diffusion in the lock-in zone.
+ Severinghaus (2012) refers to Buizert et al. (2011) for the CFC-113 based argument, but I did not see it in Buizert et al. (2011 or 2012) and all models (with/without Deddy) have quite similar results for CFC-113 in the lock-in zone.
+ The two models which do not use a Deddy term in Buizert et al. (2012) do not fit the 14CO2 peak very well (one under estimates it, the other over estimates it) but some features of other tracers in the lock-in zone are not fitted by any of the models.
The new CSIRO model simulates 14CO2 within uncertainties without dispersion in the lock-in zone (Figure 8).
+ I do not understand the d15N2 related argument as several firn models simulate the flattening of d15N2 without a dispersion term, including Trudinger et al. (1997), see comment 17.
Overall, in my opinion the most interesting question to fully explore from an equifinality point of view is: is there any hope that a dispersion term formulated as a depth-dependent Deddy flux can be adequately quantified? If I understood well the manuscript, the answer seems to be no.

AC: The existence or otherwise of dispersion in the lock-in zone is very much an open science question at the moment. Four out of six models in the Buizert et al intercomparison study required lock-in zone dispersive mixing to avoid d15N gravitational enrichment in the lock-in zone, and physical arguments can be made to expect some degree of lock-in zone dispersion (see Buizert et al. 2012). With the CSIRO model, we can fit the observations equally well with or without it, except at DSSW20K where it seems that we get better results by including it. Rather than looking at the question of whether there is clear evidence for its existence, we note the fact that some current firn models do now include it and explore two questions related to dispersion, although in a very preliminary way: 1) can we constrain dispersion (currently, no), and 2) how much does uncertainty on model predictions increase by including the possibility of dispersion. Although in general we don't find a particular need ourselves for dispersion, due to the fact that the CSIRO firn model is used to calculate the isotopic diffusion correction for various firn and ice d13CO2 measurements, we wish to assess the full uncertainty in our calculations by including this possibility.

41 - Section 4: Discussion. The manuscript clarity would be much improved if the major elements in this section were provided earlier as conclusions of the different tests/simulations performed.
AC: Agree, we now do this.

42 - p17806 l12-l18: This statement should appear much earlier in the manuscript as it has an impact on the understanding of many results. Does not it solve the above question about dispersive mixing? How does it affect the DE08-2 melt layer representation?
AC: Agree, done.
The reviewer's questions about dispersive mixing related to the monotonic assumption is answered at comment 49 below.
The question about the melt layer was answered at question 5 above.

43 - p17806 l19-20: please remove "as has always been assumed". As I see it, multitracer diffusivity constraint in firn modelling is at a very early stage of development. Moreover, does not the sentence p17807 l23-25 recommend to do what has "always been assumed"?
AC: While multi-tracer diffusivity constraint in firn modelling continues to develop, we don't agree that it is at a very early stage of development, it has certainly been our approach for some time (see Trudinger et al. 1997, 2000, 2002). It has also been our assumption for many years that more tracers will constrain diffusivity more tightly, however we agree that this may not have been a widely held assumption. The phrase has been removed.
The sentence at p17807 l23-25 is about the depth resolution of the measurements, that there is value in finer resolution.

44 - p17807 l26-l29: Then is convective mixing down to below the lock-in depth really needed to adequately match the data at DSSW20K? In my view, a negative answer to that question would be a good news for firn modelling because I see some physical contradiction between having significant convection throughout the diffusive zone and modelling the effect of gravity in the diffusive zone as at (or close to) barometric equilibrium.
AC: We get the best overall match to DSSW20K observations when we allow the convective mixing to extend into the lock-in zone, but still get an adequate match without it.

45 - p17808 l10-l11: which species could be used that is significantly different from d15N2 and 14CO2, and has a well constrained atmospheric scenario?
AC: That is a good question, to which we do not know the answer. As mentioned above, maybe comparison of d13CH4 from different southern hemisphere sites will show some insight. A recent study shows that this is no easy task (Sapart et al. 2012, ACPD special issue on firn air).

46 - p17808 l15-l17: Is the uncertainty on d13CO2 really quantifiable in the Dmolecular+Deddy frame?
AC: It is currently believed by several firn modellers (Severinghaus et al 2010, Buizert et al 2012), that there may be dispersive mixing in the lock-in zone, so to ignore this fact will lead to an underestimation of current uncertainty. As discussed above, more work is needed in this area, but at present this is a process that is included in a number of firn models, so until we can confirm or rule out its existence, it is safer to test the sensitivity of model results to it than ignore it.

47 - p17808 l24-l25: Two which extent is this due to the definition of the equifinality criterion that is based on the range of results in Buizert et al. (2012)?
AC: This is hard to say. As mentioned at point 27 above, we don't see the same pattern of spread in tracers as Buizert et al, so there was no guarantee that we would get the same range in spectral width.

48 - p17808 l28 - p17809 l1: I do not understand how these tests really constrain (quantify) isotopic fractionation.
AC: The part of the sentence in brackets "(that quantifies isotopic fractionation)" refers to Diagnostic 1, not the tests. The sentence has been rewritten to be clearer.

49 - p17809 l10-l13 and Section 5: From the above results, convective and dispersive mixing seem at least very difficult to properly constrain. For example the Deddy(z) values cannot be properly bounded if they are not monotonous (see p17806 l12-18), which is the case if significant dispersive mixing occurs. Is firn modelling just hopeless? Should the overall structure of firn models be completely revised? Is the simple parameterisation with a Deddy term of the complex physical phenomena of convective and dispersive mixing inadequate for firn modelling due to e.g. the non monotonous shape of Deddy(z)? How important are quantifiable convective and dispersive mixing compared to other transport processes in firn?
AC: We have used a monotonic assumption for molecular diffusion, however it is probably not appropriate for dispersive mixing, so some other type of smoothness constraint will be needed, such as that used in Rommelaere et al 1997. For example, including a simple dispersive mixing term that is constant with depth greatly improves the results obtained with the CIC model.
Clearly firn modelling is not hopeless, as much has been achieved to date in calibrating firn models to reproduce measurements at many sites, and for use in reconstructing the atmospheric histories of other gases. We already do well at matching the firn profiles of many gases and understand the dominant processes. The uncertainties we are looking at here, while important and interesting, are small compared to what we do know, therefore the structure of firn models is likely to develop incrementally, as more is learned about processes in the firn that are not well understood (e.g. bubble trapping fractionation, dispersion in the lock-in zone).

We should provide plausible bounds on the depth variation of Deddy from our best knowledge of the physics (as already discussed, dispersion in the lock-in zone is only a recent idea, and needs further testing), then in our modelling find ways to generate solutions that span the range of plausible solutions in order to determine uncertainties on model predictions that reflect our best knowledge of the system.
It is likely that the exponential form for Deddy to model convective mixing is too simple, and we had trouble with it at DSSW20K. We don't believe we have yet tried a suitable form for specifying dispersion.
It is convenient for us to infer a monotonic function of depth, as that guarantees smoothness, but if that is not a good approximation for reality then we must find another way to ensure smoothness.
It is important to quantify the relative impact of any uncertain processes (such as dispersion) on model predictions compared to the impact of the processes we believe we know well, and this is what we have tried to do here.

Comments on the Supplement

50 - p1 l14-15: t' is also used with fixed coordinates (z) at least in eqs. (26), (45), (46), (49) please check.

AC: This is correct. We have added "We will express some equations in terms of (t', z) - this allows easier comparison between equations in fixed and moving coordinates, and is more convenient for implementation".

51 - Equation (10): should not a total derivative be used in the second term?

AC: No, we have used t and t' to distinguish between time derivatives in fixed and moving coordinates, as already mentioned. We now include this point specifically to aid clarity.

52 - Equations (17) and (18) could be more simply introduced as a direct consequence of Equation (12), Equation (16) could be omitted.

AC: Agree, done.

53 - Figure 1: most of this complex figure is not commented, could you remove the uncommented parts or comment more? For example, the effect of compression (l131) could be illustrated by the negative values of db/dz in Fig 1c. I would be much more interested by a plot comparing the DE08 data based and Goujon et al. (2003) equation based porosities.

AC: We have improved the description of some panels in the legend of Figure 1, and added an additional comment referring to the figure. We do not wish to remove parts of the figure as we believe all parts are informative. We have not added a plot comparing DE08 data based and Goujon equation based porosities.

54 - Equation (25) seems to be a direct consequence of Equations (3) and (19a) in Rommelaere et al. (1997). Section 1.3 could be shortened, why are all equations presented in Eulerian coordinates?

AC: While we agree that equation 25 can be derived from equations in Rommelaere et al (and we did make reference to that paper), we prefer not to shorten this section. Our notation differs from that in Rommelaere, and for clarity we believe it is worth including the section as it is. Some of the equations are needed later.
We solve the Eulerian equations for *w*, then calculate *u* from *w* and *v*. We found this easier than solving the moving coordinate equation for *u* at the boundary where f goes to zero, and the equations are equivalent. We now say this in the paper.

55 - I do not understand Equation (26), which uses t' (moving coordinates) and z (fixed coordinates). How is air conservation determined?
AC: The air conservation equation comes from Rommelaere et al (1997). We have added in the '*y*' (moving coordinate) form, and another reference to Rommelaere et al.

56 - p3 l205-l206: Do you use the Goujon et al. (2003) parameterisation, or its modified version to specify the full close-off depth (f=0) from Buizert et al. (2012)?
AC: We use the parameterisation exactly as given in Buizert et al (2012). We already explained this in the main paper, but we now mention this in the supplement too.

57 - Equation (30): why does eddy diffusion (or dispersion) change the mass of the tracer into the mass of air? (see also comment 17). Why referring to Severinghaus, pers. com. (2011) rather than Severinghaus and Battle (2006) ? Are the formulations in the CSIRO model and Severinghaus and Battle (2006) equivalent or different?
AC: Eddy diffusion does not change the mass of the tracer into the mass of air. The parameterisation is given as the difference from the hydrostatic gradient. As far as we are aware, derivation of this equation was not given in Severinghaus and Battle (2006), but we are aware of it from personal communication with J. Severinghaus, thus the pers. comm. rather than reference to a paper. The final formulations are equivalent to those given in Severinghaus and Battle (2006).

58 - Extending all terms in Eq. (32), should not the last term in Eq. (33) be zero?
AC: No. It comes from the other term in the product rule on the left side of Eq. (32). We now mention this.

59 - I did not understand how Eq. (40) is obtained from Eq. (39)
AC: We now clarify this in the paper.

60 - l300-l302: then why Eq. (25) determining the air speed as a function of trapping is used?
AC: Under the commonly applied assumption that bubbles are trapped with the same pressure as currently in the firn, and without fractionation due to trapping, trapping does not affect the tracer MOLE FRACTION, but does affect tracer concentration and air conservation, and also air speed.

61 - l302-l306: I do not understand this sentence. Is J_MRx (Eq. 42) defined as a diffusive flux which does not include advection and trapping? Does not advection (u_tilde) affect the mixing ratio time derivative in eq. (43), which includes the gravitational field?
AC: Advection *u* does appear in eqn 43, and therefore does affect mixing ratio. The point we were trying to make was that advection *u* appears throughout eqn 33 for concentration but only in the last term of eqn 41 for mixing ratio, because mixing ratio does not change simply due to advection of an air parcel in a barometric gradient (tracer concentration and air concentration do though). As this point is likely to cause confusion, we have removed it.

62 - l323-l327: Thus is the upward flux of air finally negligible?
AC: No. See previous comment.

63 - Can you precise how Eq. (48) is obtained from Eq. (47)?

AC: We now describe this in the paper.

64 - The new CSIRO model (Eqs. 49-51) appears to be Eulerian (z coordinate) rather than Lagrangian (y coordinate), true? How does it affect the representation of a melt layer? Could not the model description be simplified a lot by expressing all equations only in terms of z?

AC: The CSIRO model is Lagrangian, however it is more convenient to write the equations derived in y-coordinates in terms of derivatives with respect to physical depth z, both for comparison with the equations in a fixed coordinate system, and because ice properties and concentration measurements are specified as a function of depth z. Care is needed to evaluate Δz in the finite difference equations as already described. The model description would not be simplified by deriving equations only in terms of z.

65 - l408-409: and Rommelaere et al. (1997)?

We have added a reference to Rommelaere et al (1997), but in the next sentence (line 411) instead of here, as their method is the same as in the new CSIRO model, but not the same as in the old CSIRO model.

66 - Supplementary Figure 4:

- Several dark bands are visible for some species (several preferred values of the diffusion coefficients ?), could you comment?

AC: This was mentioned at lines 502-504 on page 8 of the Supplement for the similar Figure 3 – "The clustering of points into horizontal lines in Figure 3 is due to the way the GA algorithm works, retaining solutions with low Phi and mutating or breeding them". We now remove the reference to Fig 3 so as this comment refers to both Fig 3 and 4.

- Why not show the results obtained with "real data" rather than synthetic B?

AC: Because we know the "true" diffusion coefficients for Synthetic B, so we can judge whether the calculation is doing a good job. The results for real NEEM data look similar to Synthetic B, with a clear preference for a particular value of gammaX and the solutions are dense like Synthetic B, not sparse like Synthetic A, perhaps confirming the existence of correlated errors in the data.

Technical corrections:

p17776 l2: "with some kind of automated calibration method" The method by Rommelaere et al. (1997) could be more elegantly introduced as e.g. a conjugate gradient algorithm.

We had intended "(e.g. Rommelaere et al., 1997)" to give only one example of the use of an automated calibration method for firn diffusivity. Trudinger et al. (2002) is another example, but this uses a different automatic method. We don't wish to mention specific methods like conjugate gradient here, just the general principle of automated calibration. We have changed the text to "with an automated calibration method" and added the second example.

p17777 l25, etc.: "a calibration routine ..." replace with something like "A diffusivity identification (or optimisation) algorithm"?

AC: Changed to " a genetic algorithm (GA) has been added to optimise diffusivity and other parameters".

Remove capital letter in "Synthetic" in titles of Sections 3.1 and 3.2, etc.

AC: The convention we have used is to use a capital letter when we refer to either of the cases Synthetic A or Synthetic B, but a small letter when we discuss the synthetic cases in general. These headings are therefore consistent with the rest of the text. We will leave this up to the copy editor to decide.

p17795 l6 further away?
Changed

p17796 lines 11-16 and 24-27: these convective layer formulation related comments could be regrouped.
AC: Done

Acknowledgements: I am very surprised that the NEEM acknowledgement paragraph (see http://neem.dk/publications/) is not included.
AC: All of the NEEM data we use here have been previously published in Buizert et al 2012, so the criteria at http://neem.dk/publications/ do not strictly apply to our paper. We have added a general acknowledgement for all of the firn sites that we use.

Table 1: Drill date for the second South Pole borehole is 2001 (not 2011), right?
Yes, corrected.

Supplement l79: unclear notations, not used in the following. Remove?
AC: As noted at line 69-70, we are interested in the final coordinates (y,t'), but give the others to show how these relate to (z,t). Although they are not used in the following, we prefer to retain them to aid clarity. We have explained why they are there.

Supplement l109: the total porosity s is ...
AC: done

Supplement l515-516: This definition of the lock-in depth is inappropriate for a model using a monotonous diffusivity (and density) profile.
AC: Changed to "the depth where vertical mixing goes to zero".