We thank the reviewer for their helpful and detailed comments.

Summary
The authors describe a method how to compare satellite observations, in particular TES, to model output (NASA GISS ModelE) for tropical observations between 2006 and 2009. The basic new idea is a categorized approach that maps only retrieval results and model results that exist under the same pre-defined meteorological parameters.

In the idea of a categorized filter is good. The true value of this paper is its general potential for similar studies, therefore it should be published but it has to be made somewhat clearer in order to unlock the full potential for other applications. The manuscript however has some issues I would like to adress.

General remarks

The manuscript very often refers to "high quality" retrievals, but they are never defined. What exactly do you consider a "high quality" retrieval? On p6 l22 onwards you refer to the "sensitivity" of the retrieval.
We refer to high quality retrievals as those which pass the two, standard quality filters for the TES HDO retrievals: that the HDO retrieval quality flag be set to 1 and that the sensitivity be greater or equal to 0.5. We have made this definition more explicit at P7L5 of the revised manuscript.

The observational data stems from TES, therefore it would be helpful to have brief description of the instrument.
The paper only adresses TES or in general nadir looking instruments, that should be made clear somewhere as not everyone knows what TES is.
We felt that the instrument and retrieval description at P13831-L20 was sufficient, but have indicated at P5L7 in the revised manuscript that the retrievals are based on the TES nadir measurements, and that they have a horizontal footprint of 5.3 km by 8.5 km.

Specific remarks/questions
p6 l6: The "unit" of volume mixing ratio usually is ppm/ppb. Mole fraction is not a concentration. The unit of concentration is something related to volume (molar concentration, number concentration etc.). Which of the quantities do you mean on p6 l6?
P13832-L26: Thanks. We have replaced 'concentration' with 'amount' throughout the text.

p6 l22: Are the references also true for the particular dimensions that are used in the paper?
P13833-L16: We would expect so. The relevant truncation is that of the vectors x_a^D , x_a^H , x_D and x_H and the averaging kernels from 67 to 26 levels over the 1000 to 100 hPa range. Sensitivity tests conducted as part of Risi et al. (2012) (which used the same threshold of 0.5 for the sensitivity) showed that their results were not sensitive to this truncation. Using a lower threshold (to account for some of the sensitivity 'missing' due to truncation) would result in more low-sensitivity retrievals being included, over-representation of the prior in the analysis, and artificially high agreement between the model and satellite retrievals.

p7 l7: Why do you use 825-510 hPa as borders?
P13833-L13: This is where the TES HDO/H2O retrieval is most sensitive and where analyses of this data are usually conducted. The range corresponded to that used by Yoshimura et al. (2011), and spans the ~600 hPa level used in Berkelhammer et al. (2012) and Risi et al. (2012). This has

been made more explicit at P6L23 in the revised manuscript,

p7 l11: What is lower retrieval quality?
L13834-L6: Here, lower retrieval quality refers to the fact that fewer TES retrievals were classified as usable (passing quality flag and sensitivity >= 0.5). We have slightly rephrased the revised manuscript at P7L18 to make this more clear.

p8 l21: How large are the errors on the two quantities in the respective figures? Have you fitted the correlations including the errors of the individual quantities?
P13835-L16: That is a good point. The T_S and PW_F errors are likely small, wheras the CF errors (by virtue of being calculated from cloud optical depth) are likely greater (Eldering et al., 2008). Incidentally, this was likely the reason that increasing the number of cloud bins from 15 for the 'C' categorization to 56 for the 'C_fine' categorization gave no improvement in separating good from bad retrievals (Fig 8) which we now mention at P17L16 in the revised manuscript.

We used standard OLS regression througout, and have mentioned at P8L17 the fact that this technique does not account for errors in control variables.

p8 l22: I cannot take "low quality points" or high-temperature points from the pictures.
P13835-L17: We have added a reference back to Fig 3a at P9L7 of the revised manuscript, where the Sahara stands out as having retrieval quality of less than 30%.

p9 l11: I don't understand this sentence. Is there a global measure for scatter? Is the reader supposed to know about the scatter of TES? Please be more specific here i.e. add explanation, appendix or reference.
P 13836-L3: Here, scatter refers to the dispersion of the points around each fitted regression line, which we have clarified at P9L19.

p12 l15: Here we learn suddenly that there is a "TES simulator" in the model. Please explain how it works or give a reference to a description.
P13839-L9: We have added an earlier mention of the operator being applied within a 'simulator' at P4L16, while mentioning the COSP cloud simulator, which we hope makes our approach clear.

p14 l1-l2: How do you calculate the mean retrieval quality? What is a mean Averaging Kernel? Please explain how you average Averaging Kernels.
P13840-L19: The mean retrieval quality is the proportion of HDO retrievals in a category that were classified as high quality (retrieval quality flag =1 and sensitivity > 0.5). The mean of the averaging kernels is the matrix resulting from taking the element-by-element means of all averaging kernels falling into a given category. We have clarified this at P14L11 of the revised manuscript.

p14 Sec. 5.2: I did not get a good impression how the categories are justified. The explanations here are a little bit bottom-up because you show the effects of boundaries of categories on the Avergaging Kernels. But I am sure there were considerations to size the categories independently from their Averaging Kernel relations. Otherwise it could be assumed that you solely tuned your retrieval (which for sure was not the motivation for this study).
P13841-L10: The categories were chosen to span the range of the categorical variables in the TES data. We tried to strike a balance between capturing distinctions in retrieval quality and averaging kernel structure and using as few categories as possible, now mentioned at P15L15 in the revised manuscript. Figs 8 and 9 provide some sense of the tradeoff involved before the categorical operator was applied in the model. The 'C_fine' and 'PW_fine' categorizations were designed

with this point in mind, i.e. to understand how retrieval quality and p_D separation is affected by a larger number of bins for the same variables.

p17 Sec. 5.2: Another question in this section is how the number of remaining TES observations is related to the kind of categories you are using? What is the total number of TES observations?
The distribution of all TES observations across the categories was provided for the basic 'C' categorization in Table 3, and the percentage of these that were high quality is provided in Table 4. A further breakdown was provided for one of the 'CPW' categories in Fig 7. We hope that this gives a sense of how the TES observations are distributed for the other, larger categorizations. In total, there were 202713 retrievals, originally mentioned at P13834-L4 and in the caption of Table 3.

p20 Sec. 6.1: After reading this section I still could not answer what really causes the improvement? The "resolution" of the quantitites that are categorized is downsized. Is using categories then not just a kind of additional smoothing?
P13847-L8: Using categories will smooth the predicted retrieval quality and averaging kernel structure, but (if we are interpreting the comment correctly) this type of smoothing seems inevitable for any empirical/statistical approach such as this. The point of this section was to show how shortcomings in the relationships in Figs. 11& 13 would manifest themselves in the transformed δD fields in Fig 15.

p25 Sec. 6.1: The discussion is a little weak. Because from the colored overview plots it is actually hard to judge where there a significant differences between the figures. This could be quantified better by shown residual plots (as done for section 6.3)
We assume that the reviewer is referring to Figure 15, and that rather than use the raw ModelE δD in the difference field, we should use the TES δD retrievals. We felt that the performance of the different operators should instead be evaluated before any comparison to the TES δD retrievals, or even the δD fields from the retrieval-based operator in Fig 14. Nor should, as stated at P13845-L11, the categorical operators be evaluated according to the agreement between the retrieval-based quality and p_D fields (Fig 3) and those estimated using the categorical fields (Figs. 10 & 12) but rather on how well the relationships in Figs. 11 & 13 agree with those in Fig. 4.

With that, we feel that the corresponding discussion of Fig. 15 in 6.1 accurately captures the main differences between the simple C and PW operators, and the more complicated categorizations (which were not that different), in terms of their effects on the raw ModelE δD field.

p21 l23: What is the accuracy of TES dD? I think it should be around 3% (according to Wordens 2011 paper), so I would guess the uncertainty should be between 3 and 30 permil. Thus, discussing differences of 1 permil sounds useless to me
P13848-L19: The point in this paragraph was to gauge how errors in the relationships between control variables (in this case PW_F) and p_D translate into differences in the ModelE δD fields after applying the operator. Indeed, when the main characteristics of the relationships are captured (based on the comparison between the '$LO\tau TPW_F$' and LOCPTW) the differences appear to be small, on the order of several ‰, of the same magnitude as the 1-2 ‰ observational error estimated by Risi et al. (2011) in the zonal mean (although that was over a wider latitude band).

p24 Sec. 6.3: How were the quantities used for categorization detected in the TES product? Does TES provide a CF product? If so that should be stated somewhere (or I missed it)
P13851-L8: Of the quantities in Table 1, cloud optical depth, cloud top pressure, and surface temperature are retrieved by TES, and the precipitable water-related variables were computed

directly from the TES H2O priors. These variables were used to compute the retrieval quality and averaging kernel categorization.

The CF variable was calculated as the percentage of retrievals in a grid cell that had cloud optical depth > 0.3 (clarified at P8L11 in the revised manuscript), the ModelE / ISCCP separation between clear and cloudy skies. CF was only used as a diagnostic in identifying factors influencing retrieval quality and p_D. Cloud optical depth itself was not strongly related to either because of its highly non-normal distribution, which is why we introduced the CF variable, which indeed turned out to be important. We have touched on this at P8L21 of the revised manuscript.

p25 Sec. 7: The general importance for other species should be emphasized more strongly, the benefit for the dD fields is to my opinion very limited although maybe important.
P13853-L20: We do hope that aspects of this study will be considered by others in model-satellite comparison for other species. We are reluctant to stray too far from the topic at hand, however, and would prefer leave this aspect of the discussion as-is. We hope that by making this the last point of the paper, it has been sufficiently emphasized.

Typos
*p 5, l13: replace "modeled“ with "model“ ?*P13832-L7: Corrected, thanks.

p18, l11: remove "in“ ? P13845-L6: Corrected, thanks.

p21 l1-l2: I don‘t understand the sentence beginning with "The changes ...“. Probably it could be rephrased? P13848-L14: We have rephrased this slightly to hopefully make this sentence more clear.

p23 l7: "... we examine degreee to...“, not a sentence? P13850-L6 – Corrected, thanks.

Figures
No comments, only size could be improved.
Following the comment of the first reviewer, we have removed the ‘CPW’ panels throughout, and increased the size of those figures.