Response to RC1

I think the paper will be acceptable for publication after inclusion of additional diagnostics concerning the degree to which EnKF brings useful information of the uncertainty on the state of the observed system.

Response: We are really grateful to the referee for the careful review and constructive suggestions.

The paper is a study of non-gaussianity in ensembles produced by an Ensemble Kalman Filter implemented, in a perfect model setting, on the SPEEDY intermediate meteorological model. It is largely original, and presents results of interest, such as the fact that non-Gaussianity occurs mostly in the temperature and humidity fields, and results primarily from on-off switches in the parametrization of tropical convection. I have one major comment. The Ensemble Kalman Filter that is used has ensembles with dimension 10240. If one takes the trouble of determining ensembles with such a large dimension and uses the resources that are necessary for that, it is worth evaluating those ensembles by more than the RMS error in the ensemble means and the Gaussianity, or otherwise, of those ensembles. Although the word does not appear in the present paper, it is very generally accepted that assimilation can be stated as a problem in Bayesian estimation, viz., determine the probability distribution for the state of the observed system, conditioned by the available data. Standard Kalman Filter achieves exact Bayesianity in linear and additive Gaussian situations. I think the 10240-size ensembles obtained by the authors must also be assessed in that respect. To what extent can the ensembles be considered as defining the uncertainty on the state of the observed system? The only result presented in the paper in that respect is that the spatial distributions of the RMSE in the ensemble mean and the ensemble spread are similar (top two panels of Fig. 7). I think more should be said.

Evaluation of ensembles has been discussed at length, if not for ensemble assimilation, at least for ensemble prediction (see, *e.g.*, Gneiting *et al.*, 2007). It is not possible in general to objectively assess the Bayesian character of ensembles (although it could be, albeit at a very high computational cost, in the identical twin situation, considered by the authors, in which the probability distribution of the errors affecting the data is known). But it is possible to objectively assess, on a statistical basis, two properties of ensembles. *Reliability* (also called *calibration*) is statistical consistency between the predicted PDFs and the verifying observations (reliability implies, in particular, equality between ensemble spread and RMSE in the ensemble mean). *Resolution* (also called *accuracy*, or *sharpness*) is closeness between the predicted PDFs and the observations (the RMS error in the ensemble mean is one measure of resolution). Objective scores have been defined for evaluating the degree to which an ensemble estimation system possesses those two properties. Concerning reliability (in addition to RMS-spread consistency), the easyto-obtain *rank histogram* (Hamill, 2001) gives a simple global visualisation of the degree to which it is achieved. And the *Brier score* (see, *e.g.*, Candille and Talagrand, 2005), which decomposes into a reliability and a resolution components, can also easily be computed, as well as its generalization, the *Continuous Ranked Probability Score* (*CRPS*, Hersbach, 2000). I think it is necessary to compute at least some of those scores, and to check in particular if they take different values when computed over all ensembles, or over the non-Gaussian ones only. After all, if the latter have high reliability and resolution, that will mean that the Ensemble Kalman Filter, even if it does not necessarily achieve the (rather elusive) goal of Bayesianity, provides useful information on the uncertainty on the state of the observed system, even in non-Gaussian situations.

Response: We would thank the referee for the thoughtful comments. Following the suggestions, we added the rank histogram and CRPS for all grid points and for the grid points with non-Gaussian PDF in section 4, and added related discussions in section 5.

I have in addition a number of remarks and suggestions. I put them below in approximate order of decreasing importance for each of three items.

I. Science

1. The authors use different diagnostics for identifying non-Gaussianity, *viz.*, skewness and kurtosis, Kullback-Leibler divergence, as well as the local outlier factor (LOD) for identifying outliers. Since they use ensembles with size 10240, a basic information would be the numerical values obtained for those diagnostics with a Gaussian ensemble with that size. For instance, what is the value of the quantity *DKL* (Eq. 3) for such an ensemble? And how often (if ever) does one find outliers in Gaussian samples with the LOD method as it is implemented in the paper? The only information given for Gaussian ensembles consists of Eqs (10-13) together with the associated Fig. 15. That information should come with a more precise comparison of the values obtained for Gaussian samples with the values obtained for the EnKF ensembles.

Response: We added the following description for the statistical information of KL divergence and LOF method (ll. 156-163): "The statistics of the KL divergence, SD and LOF methods with 10240 samples are evaluated numerically with 1 million trials of 10240 random draws from the standard normal distribution by the Box-Muller's method (Box and Muller 1958). The results show that the expected value of KL divergence D_{KL} is 0.0025, and its standard deviation is 0.00048. As for outlier detections, 5767 and 16088 trials have at least one outlier for SD and LOF methods, respectively. Namely, the probabilities to detect at least one outlier at a grid point are 0.58 % for the SD method and 1.6 % for the LOF method. Here, the threshold for the SD method is $\pm 5\sigma$. For the LOF method, the threshold is 8.0 and k = 20." Following the suggestion from the editor, we removed the descriptions including Eqs. 10-13 and Fig. 15. 2. Ll. 293-294, *The genesis of non-Gaussianity is explained by the convective instability*. Evidence for that is presented in the paper concerning the tropics, but not the storm tracks.

Response: We added results from the genesis of non-Gaussianity over the storm track regions in section 4 and added discussions in section 5.

3. Ll. 311-312, *In the extratropics, (the) non-Gaussianity is generally weak and seldom appears except in the storm tracks, for which there are two possible explanations.* Well, I understand the text that follows as intended at explaining why non-Gaussianity occurs rarely in the extratropics, but not why it appears more frequently in the storm tracks.

Response: In response to the previous comment, we added descriptions about the genesis of non-Gaussian PDF in the storm tracks. Therefore, this paragraph was removed.

4. Maybe I miss something, but Figure 10, and the associated text, make no sense to me. In particular, how can non-Gaussianity be identified on the Figure?

Response: We added values of KL divergence for convective instability $d\theta_e$ and spatial distributions of KL divergence for background temperature at the fourth model level at 0600 UTC and 1200 UTC on 9 February to Fig. 10.

5. Fig. 17. Concerning the SPEEDY and NICAM assimilations, a major difference is that NICAM assimilated real observations, so that model errors are present. That is to be mentioned. Response: The following phrase was added (l. 426): "although we should account for the model errors of NICAM."

6. L. 355-356, *The small cluster generated through physical processes has some physical significance*. What is the evidence that the small cluster is generated through physical processes ? I would rather suggest *The small cluster may be generated through physical processes, and have thus physical significance*.

Response: The sentence was revised accordingly (ll. 452-453): "The small cluster may be generated through physical processes and have physical significance; this should not be treated as outliers."

7. L1. 98-100, Using $\pm 3\sigma$ and $\pm 4\sigma$ thresholds, the outliers appear too frequently because 100% and 65% of all grid points statistically have at least one outlier under the Gaussian PDF. That sentence is ambiguous (and seems in contradiction with the previous one). Do you mean that, among all grid points, there is always at least one that has a $\pm 3\sigma$ outlier, and that there is a 65%

probability that there is at least one that has a $\pm 4\sigma$ outlier? Or what? And how many grid points do you consider (number of horizontal grid points *x* number of vertical levels *x* number of physical variables?). Do you consider here only univariate PDFs, or also multivariate ones?

Response: The sentences were revised accordingly (ll. 115-120): "Namely, with the threshold of $\pm 3\sigma$, we would expect to detect 27.6 outliers at every grid point. Using $\pm 4\sigma$ and $\pm 5\sigma$ thresholds, the probabilities to detect at least one outlier at a given grid point is 65 % and 0.59 %, respectively. Since the outliers appear too frequently with $\pm 3\sigma$ and $\pm 4\sigma$ thresholds, we choose the $\pm 5\sigma$ threshold for the SD method in this study." The number of grid points already have been described in section 3 (l. 170). This SD method and LOF method consider univariate PDFs, so that the following sentence was added (ll. 111-112): "Here, univariate PDFs are considered, so that SD and LOF methods are computed for each variable at each grid point separately."

8. Ll. 211-212, *The frequency of high KL divergence DKL for temperature corresponds to the time mean RMSE and DKL*. Do you mean that the frequency of high KL divergence DKL for temperature is similar to that of the time mean of RMSE and DKL, or what? And then *the pattern correlation is 0.68*. The pattern correlation between what and what exactly?

Response: The sentence was revised as follows (ll. 253-255): "The spatial distribution of frequency of high KL divergence D_{KL} for temperature is similar to that of the time mean RMSE and D_{KL} (Figs. 7 a, e, and 8 b), and the pattern correlation between the spatial distribution of mean RMSE and D_{KL} is 0.68."

9. Ll. 312-314. The authors discuss here the impact of the density of observations on the analysis. I mention for a possible future work that the density can be easily varied in the identical twin setting considered in the paper.

Response: Following the suggestion, we added discussions about the possible future work (ll. 402-409).

10. From a strict mathematical point of view, and because of sampling effects, formulæ (1-2), as well as the formula for the standard deviation σ (1. 78), are incorrect. As is well known, the denominator in the formula for σ should be *N* -1 instead of *N*. The appropriate formulæ for skewness and kurtosis are more complicated, especially if one takes into account the fact that the denominator σ in (1-2) is obtained from the same sample as the numerators. In view of the large dimension of the samples used here (10240), the error must be negligible. But have you used codes that take sampling errors into account? I suggest you mention briefly that question. And what is the exact connection between formulæ (1-2) and (10-13)? And speaking of Eq. (12), if

there is an *a priori* known bias in the sample kurtosis, why is not that bias subtracted in the first place?

Response: The skewness and kurtosis were replaced with sample skewness and sample excess kurtosis, respectively (eqs. 1, 2). The standard deviation σ was also replaced with σ =

 $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}$ (l. 87). Figures including the skewness and kurtosis were also revised. As

shown in the previous response, eqs. 10-13 were removed.

11. Ll. 281-283, *With increasing the ensemble size up to 10240, the LOFs of the small cluster and main cluster show almost the same value (Fig. 5b).* Contrary to what you seem to say here, Figures 5b and 16c are distinctly different. The small clusters have distinctly different values of LOP. That may be explained by the smaller sample in Fig. 16c (1280) than in Fig. 5b (10240), but the difference must be mentioned.

Response: We added Fig. 18d with 5120 members, and the sentences about Fig. 18 was revised accordingly (II. 331-338): "Figure 18 shows *LOF* with 80, 320, 1280, and 5120 subsamples from 10240 members for temperature at the fourth model level at the grid point B ($35.256^{\circ}N$, $146.25^{\circ}E$), as in Fig. 5b. With 80 members, there are no outliers as the *LOF* of each member is much smaller than the outlier threshold of 8. When the ensemble size is 320, four members with high *LOF* > 8 are identified as outliers. With the ensemble sizes of 1280 and 5120, 13 and 41 members construct a small cluster, respectively, but they are not outliers with the threshold of *LOF* = 8. With increasing the ensemble size up to 10240, the *LOF*s of the small cluster and main cluster show almost the same value (Fig. 5b)."

12. L. 145. It could be useful to say that the *ensemble perturbations* are the deviations from the mean of the ensemble.

Response: Revised as suggested (l. 183).

13. The authors describe in detail the local outlier factor (LOP) method (ll. 102- 125), and demonstrate it on a two-dimensional example (Fig. 3). My understanding is that it was also used in two dimensions for the diagnostics that follow (*e.g.* Fig.6). That does not seem to be said explicitly.

Response: We can apply the LOF method to multidimensional datasets (Breunig et al. 2000). Although we describe the LOF method using a simple two-dimensional example in section 2, the LOF method is applied to a one-dimensional dataset with ensemble members in section 4. For clarity, a word "two-dimensional" was added explicitly (1.121) and the following sentence was added (ll. 154-155): "Similar to the SD method, the LOF method is applied to a one-dimensional

dataset consisted of 10240 ensemble members."

14. Ll. 84-85, ... *two PDFs which are normalized by standard deviation* ... Normalization is actually not necessary for the general definition of the Kullback-Leibler divergence (that point actually does not matter here since the two PDFs that are to be compared have the same standard deviation, but what is written here may mislead an uniformed reader). Response: We agree. The phrase was removed.

15. Fig. 11. At what time (06 or 12 UTC) is $d\theta' e$ evaluated? (12 UTC, from what I understand, but say it explicitly).

Response: As mentioned in the caption of Fig. 11 and in section 4, $d\theta'_e$ is $(d\theta_{e\ 1200\ UTC} - d\theta_{e\ 0600\ UTC}) - (d\bar{\theta}_{e\ 1200\ UTC} - d\bar{\theta}_{e\ 0600\ UTC})$. We added the following phrase (II. 280-281): "The dot colors show $d\theta'_e$ evaluated from 0600 UTC to 1200 UTC 9 February,"

And change *left side* to *right side* in 1. 498 of the caption. Response: Revised as suggested (l. 612).

16. L. 338, *The number of outliers is basically one*. By which criterion for 'outlyingness'? *LOF* > 8, as indicated on l. 200? Fig. 6b does not show that there is usually one outlier. Or do you mean that (again by that particular, largely arbitrary, criterion), you observe only one outlier much more frequently than several. Although that statement is in my mind of minor interest, be more explicit, and do not wait for the conclusion to mention it.

Response: The LOF method is applied at every grid point. The sentence was revised as follows (II. 432-433): "When the outliers appear, the number of outliers is basically one per grid point,"

2. Editing

17. Ll. 332-333 ... one of three horizontal wind components ... What do you mean? See also ll. 525-526.

Response: The following sentence was added (ll. 423-424): "Here, the horizontal wind components are decomposed into three components by an orthogonal basis fixed to the earth (Satoh et al. 2008)."

18. Ll. 106-107. I would suggest to put parentheses ... number of objects (except for the object p itself) within ...

Response: Revised as suggested (ll. 125-126).

19. L. 241, The more outside members ... The formulation is awkward. I suggest The members having the largest (smallest) temperature values at 1200 UTC correspond to very large (very small) values of stability (dark red and blue points respectively)

Response: The sentence was revised accordingly (ll. 285-286): "The members with larger (smaller) temperature values at 1200 UTC correspond to larger (smaller) values of stability as shown by the warmer (colder) color."

20. L. 278, *as in Fig. 5b.* Do you mean you took the same grid point as in Fig. 5b ? Response: Yes. The following phrase was added (l. 332-333): "at the grid point B (35.256°N, 146.25°E)"

21. Ll. 263-264, *10240 members* (*see Fig. 4*) Response: Revised as suggested (1. 330).

22. Caption of Fig. 9, *As in Fig.* 8, Response: Revised as suggested.

3. English

23. L. 264, to discuss \rightarrow to identify Response: The sentence was revised accordingly (1. 330).

24. L. 280, ... are divided into outliers. \rightarrow ... are identified as outliers. Response: Revised as suggested (1. 335).

25. L. 21, the localization impact \rightarrow the impact of localization Response: Revised as suggested (1. 20).