Comments from Editor

In view of the referees' reports, I suggest (if you have not already started doing so) that you prepare a revised version of your paper, taking into account the referees' comments and suggestions. The referees mostly ask for clarifications and relatively minor modifications. Referees 1 and 3 however ask for additional diagnostics, such as rank histograms. I think those additional diagnostics should be easily obtained.

And I make as Editor an additional comment. Figure 15 does not seem to bring any material that is useful for the paper. So, unless you can cannect it more closely with the rest of the paper, I suggest you to remove it.

Response: We are really grateful to the editor for the constructive suggestions. Following the suggestions, we added new results related to the rank histograms and continuous ranked probability score (CRPS). Also, we removed Fig. 15 and related descriptions.

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 easy-to-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. Ll. 98-100, *Using ±3σ and ±4σ 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 ±3σ outlier, and that there is a 65%

probability that there is at least one that has a ±4σ 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 ±3σ, we would expect to detect 27.6 outliers at every grid point. Using ±4σ and ±5σ 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 ±3σ and ±4σ thresholds, we choose the ±5σ 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 (l. 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 $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$ (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 (ll. 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°N, 146.25°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 (l.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 uformed reader).

Response: We agree. The phrase was removed.

15. Fig. 11. At what time (06 or 12 UTC) is *d*θ'*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 (ll. 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 l. 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 (ll. 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 (l. 330).

22. Caption of Fig. 9, *As in Fig.* ***8***,

Response: Revised as suggested.

3. English

23. L. 264, *to discuss* → *to identify*

Response: The sentence was revised accordingly (l. 330).

24. L. 280, *… are divided into outliers.* → *… are identified as outliers.*

Response: Revised as suggested (l. 335).

25. L. 21, *the localization impact* → *the impact of localization*

Response: Revised as suggested (l. 20).

Response to RC2

This manuscript presents some fascinating and novel analysis of the distributions generated by an ensemble assimilation system and a low-order atmospheric model using ensembles of unprecedented size. It is generally well-written and presents a series of results that expose a number of novel aspects of the problem. It is a great addition to the ensemble literature and motivates the need for non-gaussian DA procedures rather than just using ever larger Kalman-class ensemble filters. The only weakness is that, in some places, the authors may be a bit too bold in extrapolating their results. What they have documented is the distributions from their system using a simple model and an LETKF. It is unclear whether the non-gaussian aspects of the resulting distributions are fundamental to the observing system and model, or if they might be very different with a more general data assimilation system. As an example, it is not clear if a non-gaussian DA system would result in more (or less) gaussian distributions if applied with the same nature run and observations. The authors do provide hints that the low-order model may not be a very good proxy by showing a few results with a more realistic model which look very different from the results for the SPEEDY model used here. Looks like lots of room for additional interesting studies in the future.

Response: We are really grateful to the referee for the careful review and constructive suggestions. Following the suggestions, we revised the manuscript to emphasize that the results were based on the SPEEDY-LETKF system in section 5, and to discuss the dependence of non-Gaussianity on observing systems (ll. 402-409).

Specific comments:

1. Line 29: “disappear ‘naturally”’ Not clear what it means to be ‘natural’. Do you just mean that they disappear after a while? Are there any places where things disappear ‘unnaturally’ in contrast?

Response: The phrase “disappear naturally” was removed.

2. Line 30: “1000 ensemble members may be necessary…” This is a tricky argument. Maybe it takes many fewer if the DA system isn’t assuming and enforcing (to some extent) gaussianity for increments. Or maybe there is a whole bunch more ’detailed’ structure out there in that case.

Response: Following the suggestion, the sentence was revised accordingly (ll. 29-33): “Sensitivity to the ensemble size suggests that approximately 1000 ensemble members be necessary in the intermediate AGCM-LETKF system to represent the detailed structures of the non-Gaussian PDF such as skewness and kurtosis; the higher-order non-Gaussian statistics are more vulnerable to the sampling errors due to a smaller ensemble size.” In addition, the sentence was added (ll. 76-78): "This study also discusses how many ensemble members are necessary to represent non-Gaussian PDF without contaminated by the sampling error, since in general higher-order non-Gaussian statistics are more vulnerable to the sampling error due to a limited ensemble size."

3. Line 34: "find the optimal initial state" Definition is too limited. Goal is to find some representation of a pdf. A subcase would be an optimal state.
Response: Following the suggestion, the sentence was revised (ll. 35-37). "Data assimilation is a statistical approach to estimate a posterior probability density function (PDF) using information of a prior PDF and observations. Based on the posterior PDF estimate, the optimal initial state is given for numerical weather prediction (NWP)"

4. Line 53: Actually Anderson shows that outliers occur in WEAKLY nonlinear situations, not in strongly nonlinear cases.
Response: Revised as suggested (l. 57).

5. Line 62: "non-Gaussianity will degrade the analysis." Compared to what?
Response: The sentence was revised as follows (ll. 67-69): "Since the Gaussian assumption makes the minimum variance estimator of the EnKF coincide with the maximum likelihood estimator, the non-Gaussian PDF may bring some negative impacts on the LETKF analysis."

6. Line 82: There is an issue with your definition of kurtosis throughout. The standard definition of kurtosis (check wolfram, Wikipedia) has a value of 3 for a normal distribution and the term beta_2 is usually used for this. The excess Kurtosis is this value minus 3 and has a value of 0 for normal distribution. You should make sure that both your symbols and definition are clear throughout.
Response: Following the suggestion, we replaced the kurtosis with sample excess kurtosis (eq. 2).

7. Line 91: "the PDF is considered to be non-Gaussian". This is a magic number here. Give some insight on why you picked it or make it clear that this will be discussed below.
Response: We added the following sentences (ll. 100-104): "The histogram with KL divergence $D_{KL} = 0.01$ looks approximately Gaussian while the other three histograms with larger D_{KL} values show significant discrepancies from the Gaussian function. The skewness and kurtosis measure the degrees of symmetry and tailedness, respectively, while the KL divergence D_{KL} is more suitable for measuring the degrees of difference between a given PDF and the fitted Gaussian function."

8. Line 125: k = 20 is another magic number. Give some insight into why you picked it.

Response: The sentences about choosing *k* were revised as follows (ll. 151-155): “Breunig et al. (2000) suggested that choosing *k* from 10 to 20 work well for most of the datasets. If we choose *k* too small, some objects deeply inside a cluster have a large *LOF*, and the LOF method does not work. In fact, using the dataset of KM16, *k* = 10 showed this problem, while *k* = 20 did not. Therefore, we chose *k* = 20 in this study. Similar to the SD method, the LOF method is applied to a one-dimensional dataset consisted of 10240 ensemble members.

9. Line 157: “No localization was applied, yielding the best analysis accuracy.” Do you know this to be the case? If so, state explicitly what localizations you tried (in the previous work I assume). It is surprising that no localization would be optimal since localization can also protect against nonlinear relations which are certainly occurring in the presence of non-gaussian marginal distributions that you are examining.

Response: We added the following phrase (ll. 196-199): “No localization was applied, yielding the best analysis accuracy as shown by KM16 who compared five 10240-member experiments with different choices of localization: step functions with 2000-km, 4000-km and 7303-km localization radii, a Gaussian function with a 7303-km localization radius, and no localization.”

10. Line 181: “extremely large” Compared to what?

Response: Following the suggestion, the sentence was revised (l. 221-222): “At grid point B, although the PDF seems to be closer to Gaussian, skewness $\beta_1^{1/2}$ and kurtosis β_2 are much larger than those at grid point A.”

11. Line 230: “their instability is mitigated in the model.” I had trouble with this sentence which seemed to say that the model generated instability and mitigated it at the same time. Just needs clarification.

Response: We apologize for our mistake. The word “unstable” was wrong and was replaced with “stable” (l. 272).

12. Line 250 and elsewhere: “propagates” I think that this may be a poor word choice for how marginal non-gaussianity is generated.

Response: Revised accordingly (ll. 295, 412).

13. Line 286: “Gaussian filters cannot produce accurate analysis when significant non-Gaussianity exists.” I am uncomfortable with this statement although I hope it is qualitatively

accurate. Still, I don’t think you have documented it. Kalman filters are still optimal in certain senses even for non-gaussian priors. You have provided no evidence that a non-gaussian filter would produce significantly different analyses in cases with significant non-gaussian priors. I think this statement should have a lot of caveats and suggest the need for more research.

Response: We agree and removed the sentence.

14. Line 288: “non-gaussianity of the atmosphere” First, the atmosphere doesn’t have a PDF, so it can’t be non-gaussian (caveat classical physics). Second, you have used such a simple model that it is difficult to say too much about more realistic models with similar observing systems, so be careful to not be too strong here.

Response: We removed “of the atmosphere”. As shown in the previous responses, we emphasized that the results were based on the SPEEDY-LETKF system in section 5.

15. Line 294: At least in some cases, “non-Gaussianity is explained by the convective instability.” However, your results from the more realistic model suggest it is not the only cause, and your study here only looks at a few points in detail so cannot provide strong conclusions. In addition, it is easy to demonstrate that any advective problem will generate non-gaussian distributions when the advecting flow is uncertain and the quantity being advected has gradients. This is certainly introducing non-gaussianity in all atmospheric models. See also line 302 which I think is too strong of a statement.

Response: We agree. The sentences were revised accordingly (ll. 373-374, 382-383): “With the SPEEDY model, the genesis of non-Gaussian PDF in the tropics is mainly associated with the convective instability.”, and “Therefore, convective instability is a key to non-Gaussianity genesis in the tropics in the SPEEDY model.” We also added a case of non-Gaussianity genesis from the advection in the extratropics in section 4.

16. Line 339: I believe these results are actually consistent with the results in Anderson (2010). The last section of that paper looks at results in a low-order atmospheric model, similar to SPEEDY but dry, and finds that outliers occur at local points (not for all state variables), that they form and then normally quickly disappear, and that sometimes multiple outliers can occur at the same point. Even for Lorenz-96, outliers do not form for all variables simultaneously. For instance, ensemble member 1 may be an outlier for state variable 2, while member 10 is an outlier for state variable 12.

Response: We removed “Anderson (2010)” from the sentence (l. 434) and added the following sentence (ll. 433-434): “Anderson (2010) also reported similar results using a low-order dry atmospheric model.”

Response to RC3

The authors have previously performed experiments with the LETKF and a mediumcomplexity AGCM. Previous studies have focused on the effect of localisation. In the present study the authors focus on the Non-Gaussian features of the distribution of certain variables.

I think this is an interesting topic, and the manuscript can indeed become a useful contribution to the community. To be fair, just the fact of having a 1024-member ensemble is a very rich source of information. The manuscript is in general quite clear and well-written.

I have the following comments with respect to the manuscript.

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

1. First of all, it is not clear to me how many of the experiments were performed in SPEEDY and how many of the experiments were performed in NICAM. The paper is mainly focused on SPEEDY and the experiments in NICAM are only mentioned in passing in the very last section of the paper. Is this because they were fewer and/or less detailed? Or were exactly the same experiments done in both models? If so, why did you choose SPEEDY?

Response: Although a single experiment with SPEEDY was used in the previous manuscript, we performed an additional experiment to investigate the sensitivity of non-Gaussianity to density of observations following the comment from another referee. We improved section 3 to clarify that an experiment without localization was chosen among five 10240-member experiments in KM16 (ll. 196-199). With NICAM, a single experiment was performed for just one week to get a rough idea of the realistic model’s behavior. We revised accordingly to clarify these points (ll. 416-420).

2. I agree with the other reviewers in the sense that more diagnostic metrics need to be considered besides RMSE of the analysis mean. These include: rank histograms, reliability metrics and scores. Also, the relationship between RMSE and spread can be quantified to check for the ’health’ of the ensemble system (this is related to the rank histograms too).

Response: We agree. We added rank histograms and continuous ranked probability score (CRPS) at all grid points and the grid points with non-Gaussian PDF in section 4, and discussed them in section 5. Also, we added the pattern correlation between the RMSE and ensemble spread in section 4.

3. I was very interested to see how clustering can occur in these models, and particularly excited about the fact that the set of outliers contained more than one element! In previous works (Amezcua et al, 2012; Anderson 2010) only one outlier was found. I wonder if you could say more about the size of these sets, and the way these outliers appear and disappear as the system

evolves. I guess this has to do with the distribution of maximum and minimum of a sample depending of both (a) the parent distribution and (b) the sample size. Is there anything you can mention about the distribution of the sample maxima?

Response: The outliers randomly appear and seldom affect the analysis accuracy in this study. With the more realistic models and real observations, we do not know how often the outliers are generated and how much they affect the analysis accuracy. Therefore, we would like to investigate them as a future work, and the following sentences were added (ll. 443-445): “These are the results from the simple SPEEDY model. It remains to be a subject of future research how the outliers behave with a more realistic model and real observations.”

4. It is natural to study skewness, kurtosis and KLD as the authors have done. This has been done in a univariate manner only. Can a multivariate analysis be performed? There are some ways to compute higher order moments for multivariate distributions (e.g. Mardia 1970). I am not asking to compute these if it is too hard to do it, but I am wondering if anything could be gained.

Response: We agree. This is the first study to evaluate the non-Gaussianity with huge ensemble size using atmospheric models, so that we evaluate the ensemble in a univariate field. The multivariate analysis would provide more information, and we would like to investigate it as a future work, and the following sentences were added (ll. 453-455): “The measures of non-Gaussianity are evaluated in the univariate field in this study. An extension to multivariate fields with multivariate analysis is remained as a subject of future research.”

5. The LOF method is slightly confusing and I thank the authors for adding the figure to explain it, but I wonder if there is any way to make it even clearer.

Response: The LOF method is complex. We added some additional descriptions of the LOF method in detail using Fig.3 in section 3 (ll. 141-146).

6. SPEEDY is a very "simple" (not in a bad way) model with very low resolution in time and space. It is not surprising, hence, that the source of non-Gaussianity is the parameterisation related to rain. Still, I think the analysis (including the two figures) is important. However, it should be emphasised this conclusion in valid from SPEEDY. Small scale physical processes can generate non-Gaussianity, but they are not represented in SPEEDY. I wonder if you can say anything about this with the results from NICAM?

Response: Following the suggestion, we emphasized that the results were based on the SPEEDY-LETKF system. In addition, we added a case of non-Gaussianity genesis from the advection in the extratropics in section 4 and discussed it in section 5. Also, we added the following sentence (ll. 426-428): “This result implies that the NICAM has various sources of non-Gaussianity such as smaller scale physical and dynamical processes with various interactions among different model variables,"

7. In the Ensemble Kalman Filter/Smoother, one can separate the sampling effect as having two parts, which can be approximated as being additive (Sacher and Bartello 2008; Amezcua and van Leeuwen, 2018). The more indirect part comes from the gain K coming from the sample. It is a nonlinear function of the sample covariance: bˆ2/(bˆ2+rˆ2) in the univariate case. I wouldl like to know more about the quality of K as the ensemble size changes. Note that this is related to the quality of sample B, but the 'convergence' to the true K may be slower do to the nonlinearity of the relationship.

Response: Kondo and Miyoshi (2016, KM16) investigated the analysis quality of SPEEDY model by changing the ensemble size and localization, and indicated that the analysis accuracy is improved as the ensemble size is increased with broader-scale localizations. Also, Miyoshi et al. (2014, MKI14) showed that the sampling errors were reduced from error correlations based on the B by increasing the ensemble size from 20 to 10240. Therefore, we focus on the non-Gaussianity in this study although it is important to investigate the qualities of B or K. The following sentences were added in section 1 (ll. 47-48, 65-67): "and found meaningful long-range error correlations. In addition, they reported that sampling errors in the error correlation were reduced by increasing the ensemble size.", and "Using the precious dataset of KM16 with 10240 ensemble members, we can make various investigations such as non-Gaussian statistics and sampling errors in the background error covariance. Here we focus on the non-Gaussian statistics in this study."

8. A very simple comment about figure 5: I think it would be easier to visualise if the y-axis had logarithmic scale.

Response: Logarithmic scale y-axis is not appropriate for LOF because we use three thresholds of LOF value = 5.0, 8.0, 11.0.

Non-Gaussian statistics in global atmospheric dynamics: a study with a 10240-member ensemble Kalman filter using an intermediate AGCM

Keiichi KONDO[1*] and Takemasa MIYOSHI[1, 2, 3]

[1]RIKEN Center for Computational Science, Kobe, Japan

[2]Department of Atmospheric and Oceanic Science, University of Maryland, College Park, Maryland, USA

[3]Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan

Correspondence to: Keiichi Kondo (Email: keiichi.kondo@riken.jp)

[*] Now at Meteorological Research Institute, Japan Meteorological Agency

Abstract.

We previously performed local ensemble transform Kalman filter (LETKF) experiments with up to 10240 ensemble members using an intermediate atmospheric general circulation model~~.~~ (AGCM). While the previous study focused on the impact of localization~~ impact~~ on the analysis accuracy, the present study focuses on the probability density functions (PDFs) represented by the 10240-member ensemble. The 10240-member ensemble can resolve the detailed structures of the PDFs and indicates that the non-Gaussian PDF is caused by multimodality and outliers. The results show that the spatial patterns of the analysis errors correspond well with the non-Gaussianity. While the outliers appear randomly, large multimodality corresponds well with large analysis error, mainly in the tropical regions and storm track regions where highly nonlinear~~ convective~~ processes appear frequently. Therefore, we further investigate the lifecycle of multimodal PDFs, and show that the multimodal PDFs are mainly generated by the on-off switch of convective parameterization ~~and disappear naturally.~~in the tropical regions and by the instability associated with advection in the storm track regions. Sensitivity to the ensemble size suggests that approximately 1000 ensemble members be necessary ~~to capture~~in the intermediate AGCM-LETKF system to represent the detailed structures of the non-Gaussian PDF~~.~~ such as skewness and kurtosis; the higher-order non-Gaussian statistics are more vulnerable to the sampling errors due to a smaller ensemble size.

1 Introduction

Data assimilation is a statistical approach to ~~find~~estimate a posterior probability density function (PDF) using information of a prior PDF and observations. Based on the posterior PDF estimate, the optimal initial state ~~in~~is given for numerical weather prediction (NWP). The ensemble Kalman filter (EnKF; Evensen 1994) is an ensemble data assimilation method based on the Kalman filter (Kalman 1960) and approximates the background error covariance matrix by an ensemble of forecasts. The EnKF can explicitly represent the ~~probability density function (PDF)~~PDF of the model state, where the ensemble size is essential because the sampling error contaminates the PDF represented by the ensemble. Although the sampling error is reduced by increasing the ensemble size, the EnKF is usually performed with a limited ensemble size up to O(100) due to the high computational cost of ensemble model runs. Recently, EnKF experiments with a large ensemble have been performed using powerful supercomputers. Miyoshi et al. (2014; hereafter MKI14) implemented a 10240-member EnKF with an intermediate atmospheric general circulation model (AGCM) known as the Simplified Parameterizations, Primitive Equation Dynamics model (SPEEDY; Molteni 2003)~~.~~, and found meaningful long-range error correlations. In addition, they reported that sampling errors in the error correlation were reduced by increasing the ensemble size. Further, Miyoshi et al. (2015) assimilated real atmospheric observations with a realistic model known as the Nonhydrostatic Icosahedral Atmospheric Model (NICAM; Tomita and Satoh 2004; Satoh et al. 2008; 2014) using an EnKF with 10240 members. Kondo and Miyoshi (2016; hereafter KM16) investigated the impact of covariance localization on the accuracy of analysis using a modified version of the MKI14 system.

MKI14 also focused on the PDF and reported strong non-Gaussianity, such as a bimodal PDF. ~~The EnKF inherently assumes the Gaussian PDF, but previous~~Previous studies investigated the impact of non-Gaussianity on the EnKF. Anderson (2010) reported that an *N*-member ensemble could contain an outlier and a cluster of *N*-1 ensemble members under ~~highly~~ nonlinear scenarios using the ensemble adjustment Kalman filter (EAKF; Anderson 2001). Anderson (2010) called this phenomenon ensemble clustering (EC), which leads to degradation of analysis accuracy. Amezcua et al. (2012) investigated EC with the ensemble transform Kalman filter (ETKF; Bishop et al. 2001) and local ensemble transform Kalman filter (LETKF; Hunt et al. 2007), and found that random rotations of the ensemble perturbations could avoid EC. Posselt and Bishop (2012) explored the non-Gaussian PDF of microphysical parameters using an idealized one-dimensional (1D) model of deep convection and showed that the non-Gaussianity of the parameter was generated by nonlinearity between the parameters and model output.

~~Due to the inherent Gaussian assumption of the EnKF, non-Gaussianity will degrade the analysis.~~Using the precious dataset of KM16 with 10240 ensemble members, we can make various investigations such as non-Gaussian statistics and sampling errors in the background error covariance. Here we focus on the non-Gaussian statistics in this study. Since the Gaussian assumption makes the minimum variance estimator of the EnKF coincide with the maximum likelihood estimator, the non-Gaussian PDF may bring some negative impacts on the LETKF analysis. KM16 showed that the improvement in the tropics was relatively small by increasing the ensemble size up to 10240, and suggested that the small improvement be related to the convectively dominated tropical dynamics.

This study aims to investigate the non-Gaussian statistics of the atmospheric dynamics in more detail ~~and use the results of KM16 to determine the relationships~~to investigate the relationship between the analysis error and the non-Gaussian PDF, as well as the behavior and lifecycle of the non-Gaussian PDF. To the best of the authors' knowledge, this is the first study investigating the non-Gaussian PDF using a 10240-member ensemble of an intermediate AGCM. This study also ~~aims to reveal~~discusses how many ensemble members are necessary to ~~capture~~represent non-Gaussian PDF without contaminated by the sampling error, since in general higher-order non-Gaussian statistics are more vulnerable to the sampling error due to a limited ensemble size. This paper is organized as follows. Section 2 describes measures for the non-Gaussian PDF. Section 3 describes experimental settings, and Section 4 presents the results. Finally, summary and discussions are provided in Section 5.

2 Non-Gaussian measures

~~Skewness~~Sample skewness $\beta_1^{1/2}$ and sample excess kurtosis β_2 are well-known parametric properties of a non-Gaussian PDF, and are defined as follows:

$$\beta_1^{1/2} \cancel{= \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^3}{\sigma^3}} = \frac{N}{(N-1)(N-2)}\frac{\sum_{i=1}^{N}(x_i-\bar{x})^3}{\sigma^3} \tag{1}$$

$$\cancel{\beta_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^4}{\sigma^4} - 3}\beta_2 \tag{2}$$

$$= \frac{N(N+1)}{(N-1)(N-2)(N-3)}\frac{\sum_{i=1}^{N}(x_i-\bar{x})^4}{\sigma^4} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

where x_i and $\bar{x}$ denote the ith ensemble member and N-member ensemble mean, respectively; ~~σ~~σ

denotes the sample standard deviation, i.e., $\sigma =$ ~~$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$;~~ $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$, and skewness $\beta_1^{1/2}$ represents the asymmetry of the PDF. Positive (negative) skewness $\beta_1^{1/2}$ corresponds to the PDF with the longer tail on the right (left) side. Positive (negative) kurtosis β_2 corresponds to the PDF with a more pointed (rounded) peak and longer (shorter) tails on both sides. When the PDF is Gaussian, both skewness $\beta_1^{1/2}$ and kurtosis β_2 ~~are both~~go to zero in the limit of infinite sample size. In addition, we also use Kullback–Leibler divergence (KL divergence, Kullback and Leibler 1951) from the Gaussian PDF. KL divergence is a direct measure of the difference between two PDFs. Let *p*(*x*) and *q*(*x*) be two PDFs~~ which are normalized by standard deviation σ~~. The KL divergence *D_{KL}* between the two PDFs is defined as

$$D_{KL} = \int p(x)\log\frac{p(x)}{q(x)}dx \tag{3}$$

Here, we obtain *p*(*x*) from the histogram based on the ensemble, and *q*(*x*) from the theoretical Gaussian function with the ensemble mean $\bar{x}$ and standard deviation ~~σ~~σ, respectively. D_{KL} measures the difference between the ensemble-based histogram and the fitted Gaussian function. Figure 1 shows examples of ensemble-based histograms and corresponding skewness $\beta_1^{1/2}$, kurtosis β_2, and KL divergence D_{KL}~~.~~ with 10240 samples. Here, the Scott's choice method (Scott 1979) is applied to decide the bin width for histograms. ~~In this study~~The histogram with KL divergence D_{KL} = 0.01 looks approximately Gaussian while the other three histograms with larger D_{KL} values show significant discrepancies from the Gaussian function. The skewness and kurtosis measure the degrees of symmetry and tailedness, respectively, while the KL divergence D_{KL} is more suitable for measuring the degrees of difference between a given PDF and the fitted Gaussian function. Based on the subjective observation of Fig. 1, hereafter, the PDF is considered to be non-Gaussian when $D_{KL} > 0.01$.

A non-Gaussian PDF can also be caused by outliers. Although detailed results are shown in Section 4, several ensemble members are detached from the main cluster; this also results in the large KL divergence D_{KL} shown in Fig. 2b. We tested two outlier detection methods: the standard deviation-based method (SD method) and the local outlier factor method (LOF method; Breunig et al. 2000). Here, univariate PDFs are considered, so that SD and LOF methods are computed for each variable at each grid point separately.

In the SD method, the ensemble members beyond a prescribed threshold in the unit of SD are defined as outliers. If we ~~have~~make 10240 ~~samples~~random draws from the Gaussian PDF, statistically 27.6, 0.65, and 0.0059 samples are expected beyond the ~~numbers of outliers detected by~~ $\pm 3\sigma$, $\pm 4\sigma$, and $\pm 5\sigma$ thresholds~~ are theoretically 27.6, 0.65, and 0.0059~~, respectively. Namely, with the threshold of $\pm 3\sigma$, we would expect to detect 27.6 outliers at every grid point. Using ±~~3~~4σ and ±~~4~~5σ 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 ~~because 100% and 65% of all grid points statistically have at least one outlier under the Gaussian PDF. Therefore, we set the threshold as ±5σ~~with $\pm 3\sigma$ and $\pm 4\sigma$ thresholds, we choose the $\pm 5\sigma$ threshold for the SD method in this study.

~~The~~Unlike the SD method, the LOF method is based on the local density, ~~and~~ not on the distance ~~as in~~from the ~~SD method~~sample mean. For a given two-dimensional dataset D, let $d(p, o)$ denote the distance between two objects $p \in D$ and $o \in D$. For any positive integer k, define *k-distance*(p)

to be the distance between the object *p* and the *k*th nearest neighbor. The *k-distance neighborhood* of *p*, or simply $N_k(p)$, is defined as the *k* nearest objects:

$$N_k(p) = \{q \in D \mid q \neq p, d(p, q) \leq k\text{-}distance(p)\} \tag{4}$$

The cardinality of $N_k(p)$, or $|N_k(p)|$, is greater than or equal to the number of objects (except for the object *p* itself) within *k-distance*(*p*). We define the *reachability distance* of *p* with respect to the object *o* as

$$reach\text{-}dist_k(p, o) = \max\{k\text{-}distance(o), d(p, o)\} \tag{5}$$

That is, if the object *p* is sufficiently distant from the object *o*, $reach\text{-}dist_k(p, o)$ is $d(p, o)$. If they are sufficiently close to each other, $reach\text{-}dist_k(p, o)$ is replaced by *k-distance*(*o*) instead of $d(p, o)$. Figure 3 shows a schematic diagram of $reach\text{-}dist_k(p, o)$ with $k = 3$. $N_k(p)$ includes o_1, o_2, o_3, and o_4, and $|N_k(p)|$ is 4. In Fig. 3 (a), $reach\text{-}dist_k(p, o_1)$ is $k\text{-}distance(o_1) = d(o_1, o_4)$ because $k\text{-}distance(o_1)$ is greater than $d(p, o_1)$. In contrast, in Fig. 3 (b), $reach\text{-}dist_k(p, o_1)$ is $d(p, o_1)$. We further define the *local reachability density* of *p*, or simply $lrd_k(p)$, as the inverse of the average of *reachability distance* of *p:*

$$lrd_k(p) = \frac{|N_k(p)|}{\sum_{o \in N_k(p)} reach\text{-}dist_k(p, o)} \tag{6}$$

Finally, the *local outlier factor* of *p*, denoted as $LOF_k(p)$, is defined as:

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|}. \tag{7}$$

Given a lower *local reachability density* of *p* and a higher *local reachability density* of *p*'s *k*-nearest neighbors, $LOF_k(p)$ becomes higher. $LOF_k(p)$ or simply *LOF* is approximately 1 for an object deep within a cluster, and *LOF* becomes larger around the edge of the cluster due to sparse objects on the far side from the cluster. ~~Although an~~To summarize, the LOF method focuses on the local densities of objects, and outliers are detected by comparing the local densities. For instance, when $k = 3$ in Fig. 3a, the local densities of the objects p and $o_{1, 2, 3, 4, 5}$ have all similar values because the *k-distance*(p) is similar to the *k-distances*($o_{1, 2, 3, 4, 5}$). Therefore, they are not identified as outliers. In contrast, in Fig. 3b the object p has a smaller local density than the other objects $o_{1, 2, 3, 4, 5}$ because *k-distance*(p) > *k-distances*($o_{1, 2, 3, 4, 5}$). Therefore, the object p has a larger *LOF* and is identified as an outlier. An object with *LOF* much larger than 1 may be categorized as an outlier, but it is not clear how to determine the threshold for outliers because the threshold also depends on the dataset. ~~In~~The threshold of *LOF* is chosen to be 8.0 in this study, and Section 4~~,—~~ shows the results with different values of the threshold and discusses why we ~~describe the threshold in detail.~~choose this value. k is a control parameter for the LOF method~~,~~ and depends on the dataset (Breunig et al. 2000). ~~Markus~~Breunig et al. (2000) suggested that choosing k from 10 to 20 work well for most of the datasets~~;~~. If we choose k too small, some objects deeply inside a cluster have a large *LOF*, and the LOF method does not work. In fact, using the dataset of KM16, $k = 10$ showed this problem, while $k = 20$ did not. Therefore, we chose $k = 20$ in this study. Similar to the SD method, the LOF method is applied to a one-dimensional dataset consisted of 10240 ensemble members.

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 $k = 20$.

3 Experimental settings

We use the 10240-member global atmospheric analysis data from an idealized LETKF experiment of KM16. That is, the experiment was performed with the SPEEDY-LETKF system (Miyoshi 2005) consisting of the SPEEDY model (Molteni 2003) and the LETKF (Hunt et al. 2007; Miyoshi and Yamane 2007). The SPEEDY model is an intermediate AGCM based on the primitive equations at T30/L7 resolution, which corresponds horizontally to 96 × 48 grid points and vertically to seven levels, and has simplified forms of physical parametrization schemes including large-scale condensation, cumulus convection (Tiedtke 1993), clouds, short- and long-wave radiation, surface fluxes, and vertical diffusion. Due to the very low computational cost, the SPEEDY model has been used in many studies on data assimilation (e.g., Miyoshi 2005; Greybush et al. 2011; Miyoshi 2011; Amezcua et al. 2012; Miyoshi and Kondo 2013; Kondo et al. 2013; MKI14; KM16).

The LETKF applies the ETKF (Bishop et al. 2001) algorithm to the local ensemble Kalman filter (LEKF; Ott et al. 2004). The LETKF can assimilate observations at every grid point independently, which is particularly advantageous in high-performance computation. In fact, Miyoshi and Yamane (2007) showed that the parallelization ratio reached 99.99% on the Japanese Earth Simulator supercomputer, and KM16 performed 10240-member SPEEDY-LETKF experiments within 5 minutes for one execution of LETKF, not including the forecast part on 4608 nodes of the Japanese K supercomputer. The LETKF is computed as follows. Let $\mathbf{X}$ (~~$\delta\mathbf{X}^f$~~$\delta\mathbf{X}$) denote an $n \times m$ matrix, ~~the~~whose columns ~~of which~~ are composed of m ensemble members (deviations from the mean of the ensemble~~ perturbations~~) with the system dimension n. The superscripts a and f denote the analysis and forecast, respectively. The analysis ensemble $\mathbf{X}^a$ is written as:

$$\mathbf{X}^a = \bar{\mathbf{x}}^f\mathbf{1} + \delta\mathbf{X}^f\left[\tilde{\mathbf{P}}^a(\mathbf{H}\delta\mathbf{X}^f)^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{y}^o - \mathbf{H}\bar{\mathbf{x}}^f)\mathbf{1} + \sqrt{m-1}\left(\tilde{\mathbf{P}}^a\right)^{1/2}\right] \tag{8}$$

[cf. Eqs. (6) and (7) of Miyoshi and Yamane 2007]. Here, $\bar{\mathbf{x}}^f$, $\boldsymbol{y}^o$, $\mathbf{H}$, and $\mathbf{R}$ denote the background ensemble mean, observations, linear observation operator, and observation error covariance matrix, respectively. $\mathbf{1}$ is an m-dimensional row vector with all elements being 1. The $m \times m$ analysis error covariance matrix $\tilde{\mathbf{P}}^a$ in the ensemble space is given as

$$\tilde{\mathbf{P}}^a = [(m-1)\mathbf{I}/\rho + (\mathbf{H}\delta\mathbf{X}^f)^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\delta\mathbf{X}^f)]^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^{\mathrm{T}} \tag{9}$$

[cf. Eqs. (3) and (9) of Miyoshi and Yamane 2007]. Here, ρ denotes the covariance inflation factor. As $\tilde{\mathbf{P}}^a$ is real symmetric, $\mathbf{U}$ is composed of the orthonormal eigenvectors, such that $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$. The diagonal matrix $\mathbf{D}$ is composed of the non-negative eigenvalues.

KM16 performed a perfect-model twin experiment for 60 days from 0000 UTC 1 January in the second year of the nature run, which was initiated at 0000 UTC 1 January from the standard atmosphere at rest (zero wind). The first year of the nature run was discarded as spin-up. To resolve detailed PDF structures, the ensemble size was fixed to 10240. No localization was applied, yielding the best analysis accuracy as shown by KM16 who compared five 10240-member experiments with different choices of localization: step functions with 2000-km, 4000-km and 7303-km localization radii, a Gaussian function with a 7303-km localization radius, and no localization. The observations for horizontal wind components (U, V), temperature (T), specific humidity (Q), and surface pressure (Ps) were simulated by adding observational errors to the nature run every 6 h at radiosonde-like locations (cf. Fig. 8, crosses) for all seven vertical levels, but the observations of specific humidity were simulated from the bottom to the fourth model level (about 500 hPa). The observational errors were generated from independent Gaussian random numbers, and the observational error standard deviations were fixed at 1.0 m s^{-1}, 1.0 K, 0.1 g kg^{-1}, and 1.0 hPa for U/V, T, Q, and Ps, respectively.

The non-Gaussian measures, skewness $\beta_1^{1/2}$, kurtosis β_2, and KL divergence D_{KL}, are calculated at each grid point for each variable. Outliers are diagnosed similarly at each grid point for each variable with the SD method and LOF method.

4 Results

Figure 4 shows the spatial distributions of the analysis absolute error, ensemble spread, background skewness $\beta_1^{1/2}$, kurtosis β_2, and KL divergence D_{KL} for temperature at the fourth model level (~500 hPa) at 0600 UTC 22 February. When the analysis absolute error is large, the background non-Gaussian measures also tend to be large, especially in the tropics. The peaks for skewness $\beta_1^{1/2}$, kurtosis β_2, and KL divergence D_{KL} correspond to each other. Although grid point A (16.7°S, 90.0°E) has a large KL divergence D_{KL} with large analysis absolute error, at grid point B (35.256°N, 146.25°E) with a large KL divergence D_{KL} the analysis absolute error is small ($<$ 0.08 K). This result shows that the large analysis error is not always associated with the strong non-Gaussianity at a specific time. The PDFs at grid points A and B are shown in Fig. 2a, b, respectively. The histogram at the grid point A is clearly a multimodal PDF with KL divergence $D_{KL} > 0.01$, and the right mode captures the truth (yellow star). At grid point B, although the PDF seems to be closer to Gaussian, skewness $\beta_1^{1/2}$ and kurtosis β_2 are ~~extremely large. Moreover~~much larger than those at grid point A. In fact, the PDF does not fit to the Gaussian function calculated by the ensemble mean and standard deviation. Zooming in on the left side of Fig. 2b shows a small cluster composed of 76 members detached from the main cluster; 74 members of the small cluster exceed -5σ and are categorized as outliers in the SD method. This small cluster causes the standard deviation to become large and results in the Gaussian function having a longer tail than the histogram. The small cluster should not be divided into outliers because the small cluster may have some physical significance. Scatter diagrams of *LOF* versus distance from ensemble mean for all ensemble members at grid points A and B are shown in Fig. 5a, b, respectively. At grid point A, *LOF* is not so large even at the edge of the cluster ($<$ 4), and the bimodal PDF does not influence *LOF*. In addition, all members are within $\pm 3\sigma$. Therefore, there are no clear outliers at grid point A. At grid point B, although most of the small cluster exceeds -5σ, the maximum *LOF* in the small cluster is still smaller than 3. This indicates that all members of the small cluster should not be outliers in the LOF method. Hereafter, we choose to use the LOF method.

As an outlier case, we pick up the grid point C (35.256°N, 112.5°W) in Fig. 4. The PDF at the grid point C fits the Gaussian function well, and the non-Gaussian measures are quite small (Fig. 2c). A member on the left edge of the scatter diagram in Fig. 5c has the largest $LOF > 8$, but the member is within $\pm 3\sigma$. As mentioned in Section 2, the threshold of *LOF* for outliers depends on the dataset. Figure 6 shows the number of outliers for thresholds of 5.0, 8.0, and 11.0 at 0600 UTC 22 February. There are too many outliers with threshold = 5.0, but in contrast, the number of outliers decreases markedly with threshold = 8.0 or 11.0. Based on the results, we adopt $LOF = 8.0$ as a threshold for outliers.

Figure 7 shows the spatial distributions of the time-mean analysis RMSE, ensemble spread, the background absolute skewness $\beta_1^{1/2}$, absolute kurtosis β_2, and KL divergence D_{KL}. As mentioned in KM16, the time-mean ensemble spread corresponds well to the RMSE, which is larger in the tropics. The pattern correlation between the RMSE and ensemble spread is 0.97. Moreover, the distributions of non-Gaussian measures are similar to each other and also correspond well to the RMSE and ensemble spread. The RMSE and non-Gaussian measures differ in that the non-Gaussianity is large in storm tracks, such as the North Pacific Ocean and the North Atlantic Ocean. This may be because the LETKF inhibits growing errors well in storm tracks regardless of the strong non-Gaussianity. To investigate the non-Gaussianity in more detail, Figs. 8 and 9 show the frequencies for high KL divergence $D_{KL} > 0.01$ and high *LOF* ~~≥~~ > 8, respectively. The frequency is defined as the ratio of non-Gaussianity appearance at every grid point during the 36-day period from 0000 UTC 25 January to 1800 UTC 1 March. The spatial distribution of frequency of high KL divergence D_{KL} for temperature ~~corresponds~~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. The non-Gaussianity is very strong for temperature, specific humidity, and surface pressure. In the tropics, the frequency reaches 80%, and especially the frequency in South America is over 95%, i.e., the non-~~Gaussianity~~Gaussian PDF appears for 34 days out of 36 days. In contrast, the non-~~Gaussianity~~Gaussian PDF for zonal wind hardly appears (Fig. 8 a), and the intensity of the non-Gaussianity is also weak (not shown). On the other hand, the outliers appear almost randomly and do not clearly depend on the region for any of the variables (Fig. 9), and most outliers disappear within only one or a few analysis steps. Moreover, there are no correlations between the frequency of outliers and analysis RMSE.

To investigate how the non-~~Gaussianity~~Gaussian PDF is generated, we plot the forecast and analysis update processes at 1.856°N, 168.7°E for 256 members chosen randomly from 10240 members from the analysis at 0000 UTC 9 February (157th analysis cycle) to the forecast at 0000 UTC 10 February (161st analysis cycle, Fig. ~~10~~10a). That is, Fig. ~~10~~10a shows the lifecycle of the non-~~Gaussianity~~Gaussian PDF. As the vertical axis, we introduce the convective instability $d\theta_e$, which is defined as a difference between equivalent potential temperature θ_e at the fourth model level (~500 hPa) and θ_e at the second model level (~850 hPa). Negative (Positive) $d\theta_e$ indicates a convectively unstable (stable) atmosphere. The non-~~Gaussianity~~Gaussian PDF appears in the background at the 159th cycle (1200 UTC 9 February), and the model forecast increases the KL divergence D_{KL} for $d\theta_e$ up to 0.154 and generates ~~the~~ obvious non-Gaussianity. The members of the upper side cluster at the

159th cycle generally become ~~unstable~~stable in the forecast step, and their instability is mitigated in the model. In contrast, most other members show enhanced instability. ~~Finally, the non-Gaussianity~~In the background temperature at the fourth model level, the KL divergence D_{KL} also increases from 0.003 to 0.299 for 6 h (Figs. 10b, c). Finally, the non-Gaussian PDF almost disappears at the 161st cycle (0000 UTC 10 February). Figure 11 shows a scatter diagram of 0600 UTC versus 1200 UTC 9 February for background temperature in the fourth model level for each member at 1.856~~-~~°N, 168~~.~~°7 E, and also shows histograms corresponding to the scatter diagrams. The PDF at 0600 UTC is almost Gaussian. However, at 1200 UTC, the bimodal structure ~~appears and the~~with KL divergence D_{KL} ~~becomes large from 0.003 to~~ = 0.299 ~~for 6 h~~appears. The dot colors show $d\theta_e'$ evaluated from 0600 UTC to 1200 UTC 9 February, namely, $d\theta_e' = (d\theta_{e\ 1200\ UTC} - d\theta_{e\ 0600\ UTC}) - (d\bar{\theta}_{e\ 1200\ UTC} - d\bar{\theta}_{e\ 0600\ UTC})$, where $\bar{\theta}_e$ indicates the equivalent potential temperature calculated from the ensemble mean. That is, a red (blue) dot shows more stability (instability) than the ensemble mean. The red and blue dots are clearly divided into the right and left side modes, respectively. Most members with mitigated (enhanced) instability move to the right (left) side mode. The ~~more outside~~ members with larger (smaller) temperature values at 1200 UTC ~~become much more stable (redder) or unstable (bluer), respectively~~correspond to larger (smaller) values of stability as shown by the warmer (colder) color. In addition, both right and left modes correspond to the opposite side modes in the specific humidity, respectively (not shown). That is, the members with higher (lower) temperature have lower (higher) humidity than the ensemble mean. The instability is driven by precipitation. Figure 12 is similar to Fig. 11, but for precipitation. The 10240 members are clearly divided into three clusters at

1200 UTC by the instability. The three clusters indicate the number of times cumulus parameterization is triggered. Most members in the right (left) cluster are red (blue) and show mitigation (enhancement) of the instability. Figure 13 is also similar to Fig. 11, but for zonal wind at the fourth model level. As shown in Fig. 8a, the non-Gaussianity of zonal wind is weak, and the bimodal structure appearing in ~~the~~ temperature and humidity seldom ~~propagates to~~affects the PDF of zonal wind. We ~~could not find any relationships~~found no relationship between the atmospheric instability and zonal wind. Therefore, the ~~non-Gaussianity~~ genesis of non-Gaussian PDF in the tropics is deeply related to precipitation process, which is driven by convective instability through cumulus parameterization~~.~~ in the SPEEDY model. As a result, the precipitation process mitigates the instability, ~~which raises the~~with rising temperature and ~~reduces the~~decreasing humidity. Similar results are generally obtained at other grid points with non-Gaussian PDF.

In the extratropics, non-Gaussian PDF is generated differently. To investigate the genesis of non-Gaussian PDF in the extratropics, we focus on a case around an extratropical cyclone over the Atlantic Ocean. A non-Gaussian PDF appears at 0600 UTC 15 February at 42.678°N, 48.75°W, and the KL divergence D_{KL} of background temperature increases from 0.003 to 0.460 (Fig. 14). Figure 15 is similar to Fig. 11, but for background specific humidity at the second model level (~850 hPa) versus precipitation at 42.678°N, 48.75°W at 0006 UTC 15 February. Trimodal PDFs appear in both specific humidity and precipitation. The three modes of specific humidity are clearly separated by the color, i.e., instability $d\theta_e'$. Namely, modes with larger humidity has colder colors (smaller $d\theta_e'$ corresponding to more instability). However, the three modes of precipitation show no clear dependence on $d\theta_e'$. Therefore, the trimodal PDF of specific humidity would not be driven by the cumulus parameterization. Next, the relationship between background specific humidity and meridional wind at the second model level (~850 hPa) is shown in Fig. 16. The members in the left mode have lower specific humidity with relatively stronger northerly wind. If we look at the fourth model level (~500 hPa) for these members with lower humidity, they have relatively weaker northerly wind and warm temperature (not shown). Namely, instabilities are mitigated by the northerly advection of dry air at the lower troposphere and by warm temperature at the mid troposphere. In this case study, the non-Gaussianity genesis in the extratropics is associated with the advections. This is only an example, and the non-Gaussianity genesis in the extratropics is generally more complicated and would be affected by not only vertical stratification but also larger-scale atmospheric phenomena such as extratropical cyclones and advections. Here, we do not go into details for different cases of non-Gaussianity genesis, but instead, this is further discussed~~ in detail~~ in Section 5.

The non-Gaussian measures are sensitive to the ensemble size due to sampling errors. Figure ~~14~~17 shows the spatial distributions of the skewness $\beta_1^{1/2}$, kurtosis β_2, and KL divergence D_{KL} for temperature at the fourth model level (~500 hPa) at 0600 UTC 22 February with 80, 320, and 1280 subsamples from 10240 members, respectively. Skewness $\beta_1^{1/2}$, kurtosis β_2, and KL divergence D_{KL} with 80 members contain high levels of contaminating errors originating from sampling errors, and the non-Gaussian measures are difficult to distinguish from the contaminating errors. With increasing the ensemble size up to 1280, the sampling errors become smaller by gradation. With 1280 members, the sampling errors are essentially removed, and the distributions are comparable to those with 10240 members (see Fig. 4). ~~Thus~~ Therefore, a sample size of about 1000 members is ~~sufficient~~necessary to ~~discuss~~represent non-~~Gaussianity~~Gaussian PDF. ~~The sampling error depends on the ensemble size, and it is known that the sampling error decreases in inverse proportion to the square root of the sample size. Pearson (1931) derived exact expressions for the relationships between the sample size and standard deviations of skewness $\beta_1^{1/2}$ and kurtosis β_2 under the distribution. The expected mean and standard deviation of distribution of skewness $\beta_1^{1/2}$ are written as~~

~~$\mu\left(\beta_1^{1/2}\right) = 0$ (10)~~

~~$\sigma\left(\beta_1^{1/2}\right) = \sqrt{\frac{6(N-2)}{(N+1)(N+3)}}$ (11)~~

~~and the expected mean and standard deviation of distribution kurtosis β_2 are given by~~

~~$\mu(\beta_2) = -\frac{6}{N+1}$ (12)~~

~~$\sigma(\beta_2) = \sqrt{\frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}}$ (13)~~

~~where N is the sample size. Based on Eqs. (10)–(13), Fig. 15 plots the expected means and standard deviations of skewness $\beta_1^{1/2}$ and kurtosis β_2 with increasing sample size up to 10240. When the ensemble size is small, it is difficult to extract the non-Gaussian signals because the sampling error is almost the same order as the non-Gaussian signals. In contrast to skewness $\beta_1^{1/2}$, the expected value of kurtosis β_2 is negative and convergence is relatively slow. Indeed, the distribution of kurtosis β_2 with 80 members is mostly covered by negative values (Fig. 14d).~~ The outliers also depend on the sample size. Figure ~~16~~18 shows *LOF* with 80, 320, ~~and~~ 1280, and 5120 subsamples from 10240 members for temperature at the fourth model level~~,~~ at the grid point B (35.256°N, 146.25°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, ~~just~~ four members with high *LOF* ~~≥~~> 8 are ~~divided into~~identified as outliers. ~~Moreover, with an~~ With the ensemble ~~size~~sizes of 1280~~, 11~~ 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).

We saw a good agreement between the RMSE and ensemble spread (Figs. 7a, b), but it is useful to further evaluate the 10240-member ensemble using ranked probability scores. The rank histogram (Hamill and Collucci 1997, Talagrand and Vautard 1997, Anderson 1996, Hamill 2001) evaluates the reliability of ensemble statistically. Figure 19 shows almost flat rank histograms at all grid points and the grid points with non-Gaussian PDF. The truth is known in this study and used as a verifying analysis. The flat rank histograms correspond to healthy background ensemble distributions. The continuous ranked probability score (CRPS, Hersbach 2000) is another method to evaluate ensemble distributions, decomposed into reliability, resolution and uncertainty as

$$\mathrm{CRPS} = \mathrm{Reli} - \mathrm{Resol} + U. \qquad (10)$$

Here, the reliability Reli becomes zero under the perfectly reliable system. The resolution Resol indicates the degree to which the ensemble distinguishes situations with different frequencies of occurrence, and is associated with the accuracy or sharpness. The uncertainty *U* measures the climatological variability. The reliability, resolution and uncertainty are given on the prescribed area as

$$\mathrm{Reli} = \sum_{i=0}^{N} \bar{g}_i (\bar{o}_i - p_i)^2 \, , \tag{11}$$

$$p_i = \frac{i}{N},$$

$$U - \mathrm{Resol} = \sum_{i=0}^{N} \bar{g}_i \bar{o}_i (1 - \bar{o}_i), \tag{12}$$

$$U = \sum_{k,l<k} w_k \, w_l |y^k - y^l|, \tag{13}$$

[cf. Eqs 36, 37 and 19 in Hersbach 2000, respectively]. Here, $\bar{g}_i$ is the area-weighted average width of the bin *i* between consecutive ensemble members x_i and x_{i+1}, and $\bar{o}_i$ is the area-weighted average frequency that the verifying analysis is less than $(x_{i+1} + x_i)/2$. *N* denotes an ensemble size. In this study, y^k and y^l indicate the anomalies between the background ensemble mean and monthly climatology computed from a 30-year nature run at the grid points *k* and *l*, respectively. The weights $w_{k,}$, w_l are proportional to the cosine of latitude. Table 1 shows that the reliability is closer to zero and that the resolution is much higher at all grid points than at the grid points with non-Gaussian PDF. Therefore, the non-Gaussian PDF has a negative impact on updating the state variables for the LETKF. The smaller uncertainty at the grid points with non-Gaussian PDF reflects generally smaller variations in the tropics where the non-Gaussian PDFs frequently appear. Similar results are obtained for the other variables.

5 Summary and discussions

~~Gaussian~~Kalman filters~~, such as EnKFs, cannot treat non-Gaussian PDF properly. That is,~~ provide the ~~Gaussian filters cannot produce accurate analysis when significant non-Gaussianity exists. Therefore, this~~minimum variance estimator, which coincides with the maximum likelihood estimator under the Gaussian assumption. This study investigated the non-~~Gaussianity of the atmosphere~~Gaussian PDF and its behavior using ~~a~~the SPEEDY-LETKF system with 10240 members. ~~The non~~Non-Gaussian ~~PDF appears~~PDFs appear frequently in the areas where the RMSE and ensemble spread are ~~large~~larger. Moreover, an ensemble size of about 1000 is ~~large enough~~necessary to ~~resolve~~represent the non-Gaussian PDF ~~and~~which is more vulnerable to ~~mitigate~~ the sampling error.

The non-Gaussian PDF appears frequently in the tropics and the storm ~~tracks~~track regions over the Pacific and Atlantic Oceans, particularly for temperature and specific humidity, but not for winds. ~~The~~With the SPEEDY model, the genesis of non-~~Gaussianity~~Gaussian PDF in the tropics is ~~explained by~~mainly associated with the convective instability. These results suggest that the non-~~Gaussianity~~Gaussian PDF be mainly driven by precipitation processes such as cumulus parameterization but much less by dynamic processes. Generally, the atmosphere in the tropics tends to become unstable, and the convective instability is mitigated by vertical convection with precipitation. In the SPEEDY model, a simplified mass-flux scheme developed by Tiedtke (1993) is applied. Convection occurs when either the specific or relative humidity exceeds a prescribed threshold (Molteni 2003). The members that hit the threshold have precipitation, and this process mitigates their own convective instability resulting in a temperature rise and humidity decrease. In contrast, the members with no or little precipitation enhance or cannot mitigate their own convective instability. ~~Therefore, these results indicate that convective instability is the key to non-Gaussianity~~

~~genesis. Moreover, in the tropics, non-Gaussianity genesis may be led by a large ensemble spread due to less observational information, which comes from sparse observations and short decorrelation lengths in the tropics. KM16 mentioned that the decorrelation length was generally short in the tropics. Less observational information results in a large ensemble spread, where many members easily reach the threshold of cumulus parameterization, and this accelerates the non-Gaussianity genesis. In addition, if we use more realistic models with advanced parameterization schemes or cloud microphysics, the process of non-Gaussianity genesis would become more complex~~Therefore, convective instability is a key to non-Gaussianity genesis in the tropics in the SPEEDY model.

~~In the extratropics, the non-Gaussianity is generally weak and seldom appears except in the storm tracks, for which there are two possible explanations. The first is the high density of observations. In contrast to the tropics, there are many observations in the extratropics, mainly over land. Many observations help improve the analysis and cause contraction of the ensemble spread. The contracted ensemble spread can inhibit generation of non-Gaussianity by reducing the number reaching the threshold of parameterization. Second, the horizontal long-range correlation of the atmosphere may be related to the non-Gaussianity. As mentioned in KM16, the long-range correlation in the extratropics has a beneficial influence on the accuracy of the analysis, particularly in sparsely observed areas, such as over the ocean. That is, there are many observations to be assimilated in sparsely observed areas through long-range correlation. This indicates that the long-range correlation plays a role in reducing the analysis error and contributes to inhibition of the non-Gaussianity genesis.~~

In the extratropics, the non-Gaussian PDF is generally weak and seldom appears except in the storm track regions, where the genesis of non-Gaussian PDF is also associated with instabilities, but with different processes from the tropics. This study focused on a case near the extratropical cyclone in the North Atlantic, and the results showed that the instability was associated with the horizontal advections. The members with their instabilities mitigated had lower humidity at the lower troposphere and higher temperature at the mid troposphere by meridional advections. In contrast, the members with higher humidity at the lower troposphere and lower temperature at the mid troposphere enhanced their instability. Moreover, the precipitation process through the cumulus parameterization did not explain the non-Gaussian PDF. Precipitation associated with extratropical cyclones is usually caused by synoptic-scale baroclinic instabilities and does not mitigate the local instability completely.

As mentioned in Section 4, to generalize the process of non-Gaussianity genesis in the extratropics is not simple. The non-Gaussianity genesis is generally associated with instability from various processes such as the convection, advection and larger-scale atmospheric phenomena, so that it is very difficult to find general mechanisms of the non-Gaussianity genesis in the extratropics even for the simple SPEEDY model. Furthermore, if we use more realistic models with complex physics schemes, the process of non-Gaussianity genesis would be much more diverse and complicated. This is partly why we did not go into details to investigate different cases of non-Gaussianity genesis with the SPEEDY model.

Although the frequency of non-Gaussian PDF seems to depend primarily on the density of observations, it also seems to reflect the contrast between the continents and oceans (see Fig. 8). To investigate the sensitivity to the spatial density of observations, we performed an additional experiment in which 333 radiosonde stations were added over the tropical oceans, the North Pacific Ocean and the North Atlantic Ocean using 10240 ensemble members. The results showed that the frequency and intensity of non-Gaussianity were almost unchanged (not shown). How does non-Gaussianity depend on the spatial and temporal densities of observations? This remains to be a subject of future research.

The non-Gaussianity is less frequent in the wind components not only ~~on~~in the time scale of 1 month but also for the snapshot, although the dynamic process of the atmosphere is a nonlinear system. Moreover, the non-~~Gaussianity seldom seems to propagate from the ~~Gaussian PDFs of temperature and specific humidity ~~fields to~~seldom affect the PDFs of the wind components. We hypothesize that the model complexity may be a reason for this. The SPEEDY model could not resolve some local interactions between wind components and other variables due to its coarse resolution and simplified processes. With more realistic models, physical processes are much more complex, and the local interactions can also be represented. Indeed, we obtained widely distributed non-Gaussianity with a 10240-member NICAM-LETKF system with 112-km horizontal resolution assimilating real observations from the National Centers for Environmental Prediction (NCEP) known as PREPBUFR from 0000 UTC 1 November to 0000 UTC 8 November (Miyoshi et al. 2015). Figure ~~17~~20 shows the spatial distributions of background KL divergence of zonal wind and temperature at the second model level (~850 hPa) for SPEEDY at 0000 UTC 1 March and one of three horizontal wind components and temperature at the fifth model level (~850 hPa) for the NICAM at 0000 UTC 8 November 2011. Here, the horizontal wind components are decomposed into three components by an orthogonal basis fixed to the earth (Satoh et al. 2008). With NICAM, the non-Gaussianity appears globally not only in the temperature field but also in the wind component~~.~~ although we should account for the model errors of NICAM. This result implies that the NICAM has various sources of non-Gaussianity such as smaller scale physical and dynamical processes with various interactions among different model variables, and suggests the limitation of this study using the SPEEDY model. In the realistic situation, we would have an abundance of non-Gaussianity.

The outliers appear almost randomly regardless of locations, levels, and variables, and the lifetime is about a few analysis steps. ~~The~~When the outliers appear, the number of outliers is basically one per grid point, but sometimes the number is more than one. Anderson (2010) also reported similar results using a low-order dry atmospheric model. These results seem not to be consistent with ~~Anderson (2010) and~~ Amezcua et al. (2012) who reported that just one outlier appeared with the ensemble square root filters in low-dimensional models and that the outlier did not rejoin the cluster easily. These properties of their outlier and our outliers in the SPEEDY model are somewhat different. In the low-dimensional models, a certain ensemble member ~~becomes~~tends to become an outlier at all grid points and all variables. In contrast, the outliers in the SPEEDY model appear at just some grid points but not all grid points and do not appear in all variables simultaneously. In addition, the negative influence of outliers on the analysis accuracy may be sufficiently small in high-dimensional models due to the randomness and short longevity of outliers. In fact, the results showed no clear correspondence between the outlier frequency and analysis accuracy. These are the results from the simple SPEEDY model. It remains to be a subject of future research how the outliers behave with a more realistic model and real observations.

As measures of non-Gaussianity, skewness, kurtosis, and KL divergence for the non-Gaussianity, and the SD and LOF methods for outliers, are introduced and compared with each other. The KL divergence is a more suitable measure because it measures the direct difference between the ensemble-based histogram and the fitted Gaussian function. The LOF method is better than the SD method because it can detect the outliers depending on the density of objects. Although it is easy to detect the outliers using the SD method, misdetection of outliers is possible because this method categorizes a small cluster far from the main cluster into outliers. The small cluster may be generated through physical processes ~~has some~~and have physical significance~~ and~~; this should not be ~~divided into~~treated as outliers. The measures of non-Gaussianity are evaluated in the univariate field in this study. An extension to multivariate fields with multivariate analysis is remained as a subject of future research.

~~The non~~Non-Gaussian measures ~~are~~tend to be more sensitive to the sampling error due to the limited ensemble size (see Figs. ~~14, 16~~17, 18). When the ensemble size is small, it is difficult to determine whether a split member is a real outlier or a sample from a small cluster. Amezcua et al. (2012) discussed the outliers by skewness using the 20-member SPEEDY-LETKF and reported that the skewness is clearly large in the tropics and the Southern Hemisphere for the temperature and humidity fields. These results were not consistent with those of the present study because the outliers appear randomly. However, this inconsistency may have been due to the small ensemble size. The large skewness of Amezcua et al. (2012) could possibly indicate the non-Gaussianity rather than the outliers with a large ensemble size. Having a sufficient ensemble size, suggested to be about 1000 according to this study, would be essential when discussing about non-Gaussianity and outliers.

Data availability

All data and source code are archived in RIKEN Center for Computational Science and are available upon request from the corresponding authors under the license of the original providers. The original source code of the SPEEDY-LETKF is available at https://github.com/takemasa-miyoshi/letkf.

Acknowledgments

~~We~~The authors are grateful to the members of the Data Assimilation Research Team, RIKEN R-CCS and of the Meteorological Research Institute for fruitful discussions. The SPEEDY-LETKF code is publicly available at https://github.com/takemasa-miyoshi/letkf~~http://code.google.com/p/miyoshi/~~. Part of the results was obtained using the K computer at the RIKEN R-CCS through proposal numbers ra000015 and hp150019. This study was partly supported by JST CREST Grant number JPMJCR1312, and JSPS KAKENHI Grant number JP16K17806.

References

Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, J. Climate, 9, 1518–1530, 1996.

Anderson, J. L.: An ensemble adjustment Kalman filter for data assimilation, Mon. Wea. Rev., 129, 2884-2903, 2001.

Anderson, J. L.: A non-Gaussian ensemble filter update for data assimilation, Mon. Wea. Rev., 138, 4186-4198, 2010.

Amezcua, J., Ide, K., Bishop, C. H., and Kalnay, E.: Ensemble clustering in deterministic ensemble Kalman filters, Tellus, 64A, 1-12, 2012.

Bishop, C. H., Etherton, B. J. and Majumdar, S. J.: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. Mon. Wea. Rev., 129, 420-436, 2001.

Box, G. E. P. and Muller, Mervin E.: A note on the generation of random normal deviates, Ann. Math. Statist., 29, 610-611, doi:10.1214/aoms/1177706645.

Breunig, M. M, Kriegel,H. P. R., Ng, T., and Sander, J.: LOF: Identifying density-based local outliers, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 93-104, ~~dio~~doi: 10.1145/335191.335388, 2000.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res. 99C5, 10143-10162, 1994.

Greybush, S. J., Kalnay, E., Miyoshi, T., Ide, K., and Hunt, B. R.: Balance and ensemble Kalman filter localization techniques, Mon. Wea. Rev., 139, 511-522, 2011.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Wea. Rev., 129, 550-560, 2001.

Hamill, T., and Colucci, S. J.: Verification of Eta–RSM short- range ensemble forecasts, Mon. Wea. Rev., 125, 1312–1327, 1997.

Hersbach, H.: Decomposition on the continuous ranked prob- ability score for ensemble prediction systems, Wea. Forecasting, 15, 559–570, 2000.

Hunt, B. R., Kostelich, E. J., and Syzunogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D, 230, 112-126, 2007.

Kalman, R. E.: A new approach to linear filtering and predicted problems, J. Basic Eng.. 82, 35-45, 1960.

Kondo, K. and Miyoshi, T.: Impact of removing covariance localization in an ensemble Kalman filter: experiments with 10240 members using an intermediate AGCM, Mon. Wea. Rev., 144, 4849-4865, 2016.

Kondo, K. and Miyoshi, T., and Tanaka, H. L.: Parameter sensitivities of the dual-localization approach in the local ensemble transform Kalman filter, SOLA, 9, 174-178, 2013.

Kullback, S., and Leibler, R. A.: On information and sufficiency, The Annals of Mathematical Statistics, 22, 79-86, 1951.

Miyoshi, T.: *Ensemble Kalman Filter Experiments with a Primitive-equation Global Model*. PhD Thesis, University of Maryland, College Park, 226 pp., 2005.

Miyoshi, T.: The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter, Mon. Wea. Rev., 139, 1519–1535, doi: 10.1175/2010MWR3570.1, 2011.

Miyoshi, T. and Yamane, S.: Local ensemble transform Kalman filtering with an AGCM at a T159/L48 resolution, Mon. Wea. Rev., 135, 2841-3861, 2007.

Miyoshi, T. and Kondo, K.: A multi-scale localization approach to an ensemble Kalman filter, SOLA, 9, 170-173, 2013.

Miyoshi, T., Kondo, K., and Imamura, T.: 10240-member ensemble Kalman filtering with an intermediate AGCM. Geophys. Res. Lett., 41, 5264–5271, doi: 10.1002/2014GL060863, 2014.

Miyoshi, T., Kondo, K., and Terasaki, K.: Big Ensemble Data Assimilation in Numerical Weather Prediction, Computer, 48, 15-21, doi:10.1109/MC.2015.332, 2015.

Molteni, F.: Atmospheric simulations using a GCM with simplified physical parameterizations. I: model climatology and variability in multi-decadal experiments, Clim. Dyn., 20, 175-191, 2003.

Ott, E., and Coauthors: A local ensemble Kalman filter for atmospheric data assimilation, Tellus, 56A, 415–428, 2004.

Pearson, Egon S.: Note on tests for normality, Biometrika, 22, 423-424, 1931.

Posselt, D., and Bishop, C. H.: Nonlinear parameter estimation: comparison of an ensemble Kalman smoother with a Markov chain Monte Carlo algorithm, Mon. Wea. Rev., 140, 1957-1974, 2012.

Satoh, M., Matsuno, T., Tomita, H., Miura, H., Nasuno, T., and Iga, S.: Nonhydrostatic icosahedral atmospheric model (NICAM) for global cloud resolving simulations, Journal of Computational

Physics, the special issue on Predicting Weather, Climate and Extreme events, 227, 3486-3514, doi:10.1016/j.jcp.2007.02.006, 2008.

Satoh, M., Tomita, H., Yashiro, H., Miura, H., Kodama, C., Seiki, T., Noda, A. T., Yamada, Y., Goto, D., Sawada, M., Miyoshi, T., Niwa, Y., Hara, M., Ohno, T., Iga, S., Arakawa, T., Inoue, T., and Kubokawa, H.: The non-hydrostatic icosahedral atmospheric model: Description and development, Progress in Earth and Planetary Science, 1, 18, doi:10.1186/s40645-014-0018-1, 2014.

Scott, D. W.: On optimal and data-based histograms, Biometrika, 66, 605-610, doi:10.1093/biomet/66.3.605, 1979.

Talagrand, O., and Vautard, R.: Evaluation of probabilistic pre-diction systems. Proc. ECMWF Workshop on Predictability, Reading, United Kingdom, ECMWF, 1–25, 1997.

Tiedtke, M: A comprehensive mass flux scheme for cumulus parameterization in large-scale models, Mon. Wea. Rev., 117, 1779-1800, 1993.

Tomita, H., and Satoh, M.: A new dynamical framework of nonhydrostatic global model using the icosahedral grid, Fluid Dyn. Res., 34, 357-400, 2004.

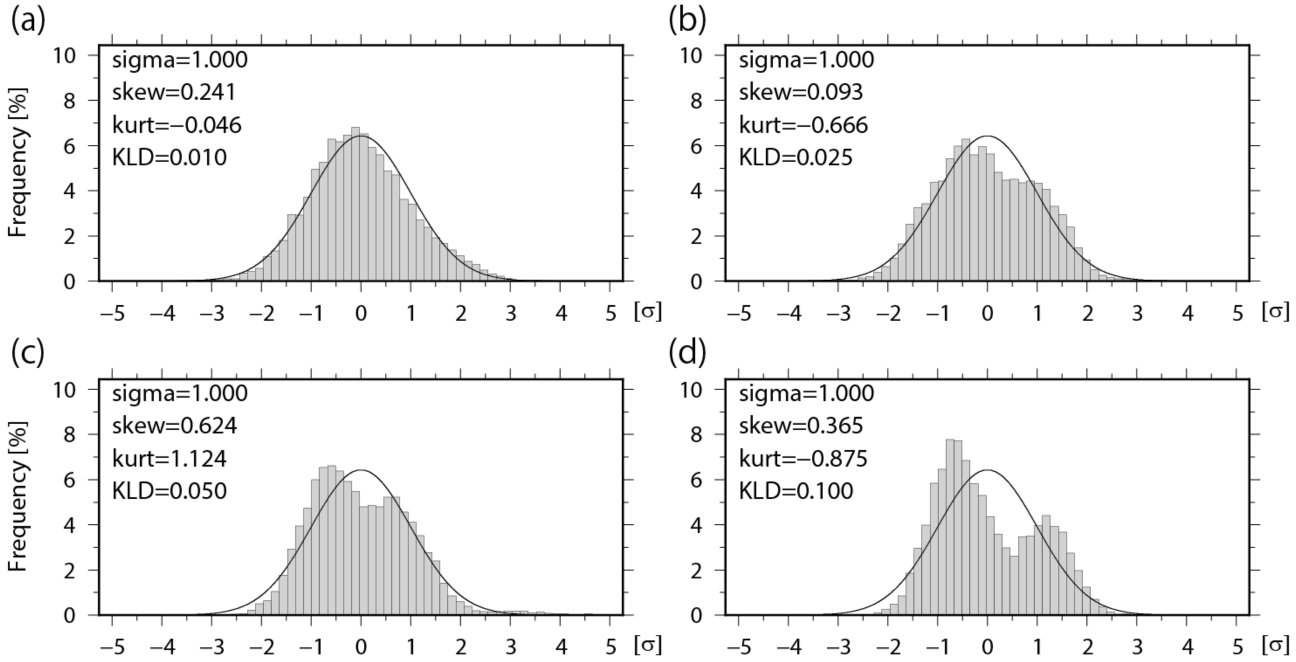


Figure 1: Ensemble-based histograms with 10240 ensemble members when the Kullback–Leibler (KL) divergence D_{KL} = (a) 0.010, (b) 0.025, (c) 0.050, and (d) 0.100. Solid lines indicate fitted Gaussian functions. Skewness (skew) and kurtosis (kurt) are also shown in the figure.

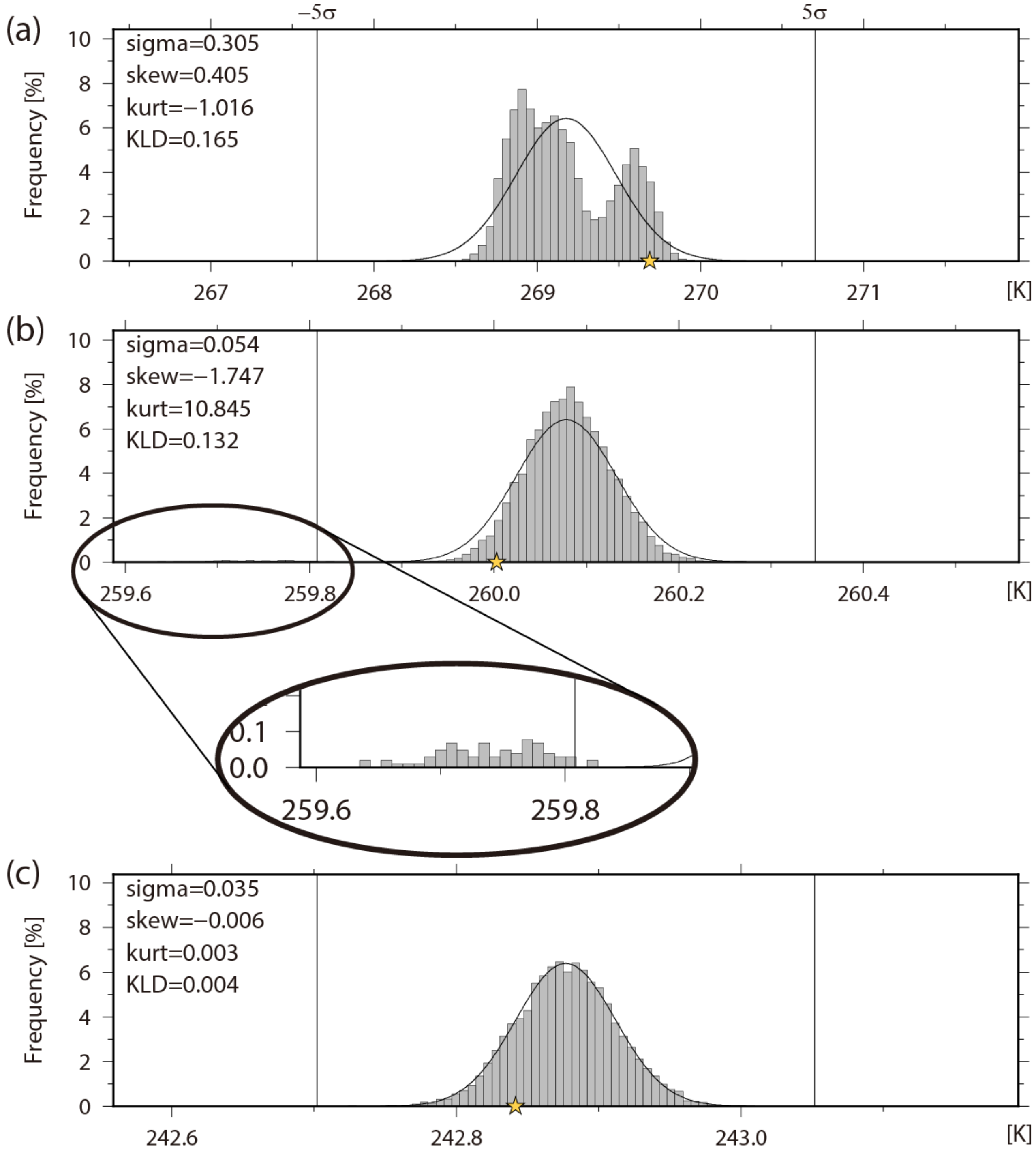


Figure 2: Histograms of background temperature (K) at the fourth model level (~500 hPa) at (a) grid point A (16.7°S, 90.0°E), (b) grid point B (35.256°N, 146.25°E), and (c) grid point C (35.256°N, 112.5°W). The ~~orange~~yellow star shows the truth.

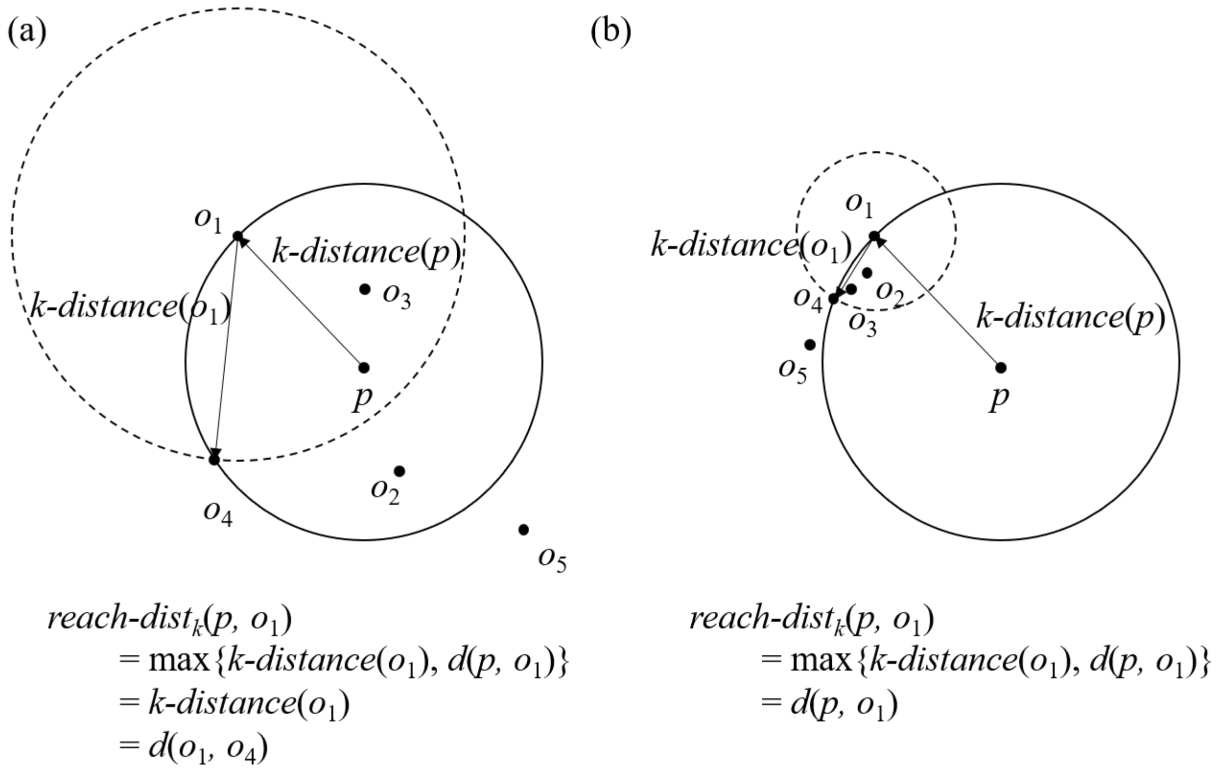

(a)
o_1
k-distance(p)
o_3
k-distance(o_1)
p
o_4
o_2
o_5
$reach\text{-}dist_k(p, o_1)$
$= \max\{k\text{-}distance(o_1), d(p, o_1)\}$
$= k\text{-}distance(o_1)$
$= d(o_1, o_4)$
(b)
o_1
k-distance(o_1)
o_4
o_2
o_3
k-distance(p)
o_5
p
$reach\text{-}dist_k(p, o_1)$
$= \max\{k\text{-}distance(o_1), d(p, o_1)\}$
$= d(p, o_1)$

Figure 3: Schematic diagrams of *reach-dist*$_k$ (p, o) with $k = 3$ for (a) uniformly distributed data and (b) data with an asymmetrical distribution.

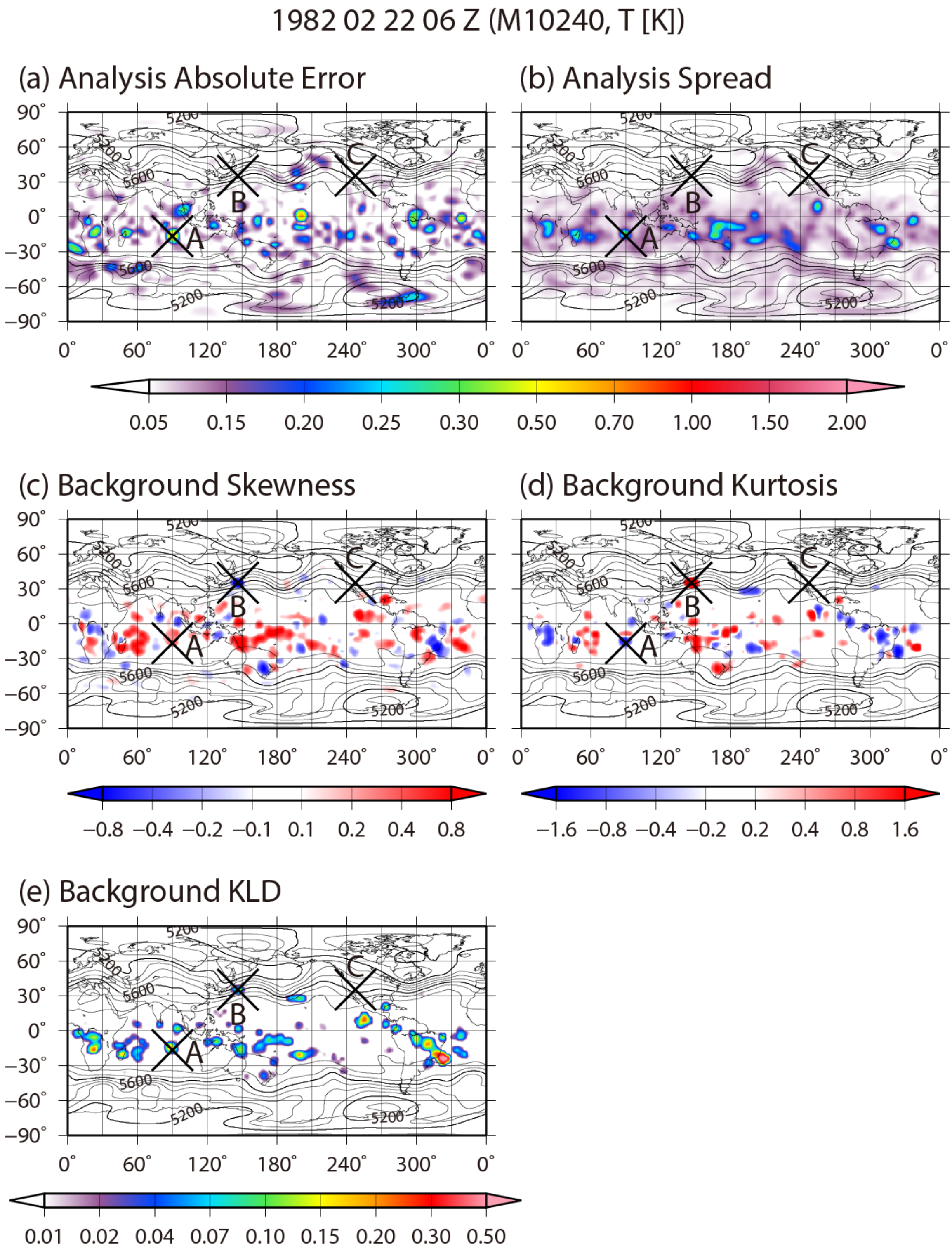


Figure 4: Spatial distributions of (a) analysis absolute error, (b) analysis ensemble spread, (c) background skewness, (d) background kurtosis, and (e) background KL divergence for temperature at the fourth model level (~500 hPa) ~~on~~at 0600 UTC 22 February. Contours indicate geopotential height of the ensemble mean at the 500 hPa level.

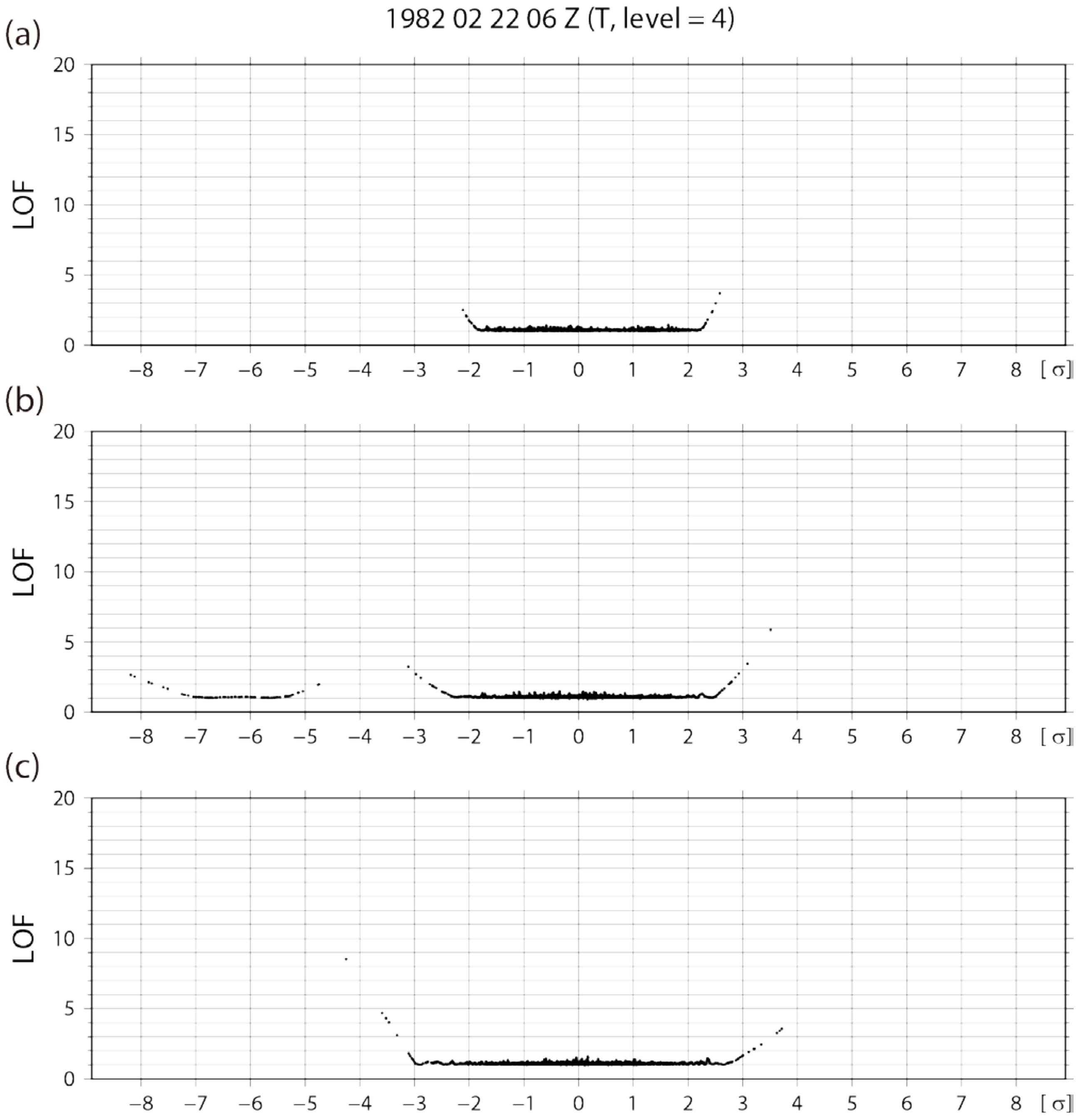


Figure 5: Scatter diagrams of the local outlier factor method (*LOF*) versus distance from the ensemble mean for all ensemble members for background temperature at the fourth model level (~500 hPa) at (a) grid point A (16.7°S, 90.0°E), (b) grid point B (35.256°N, 146.25°E), and (c) grid point C (35.256°N, 112.5°W).

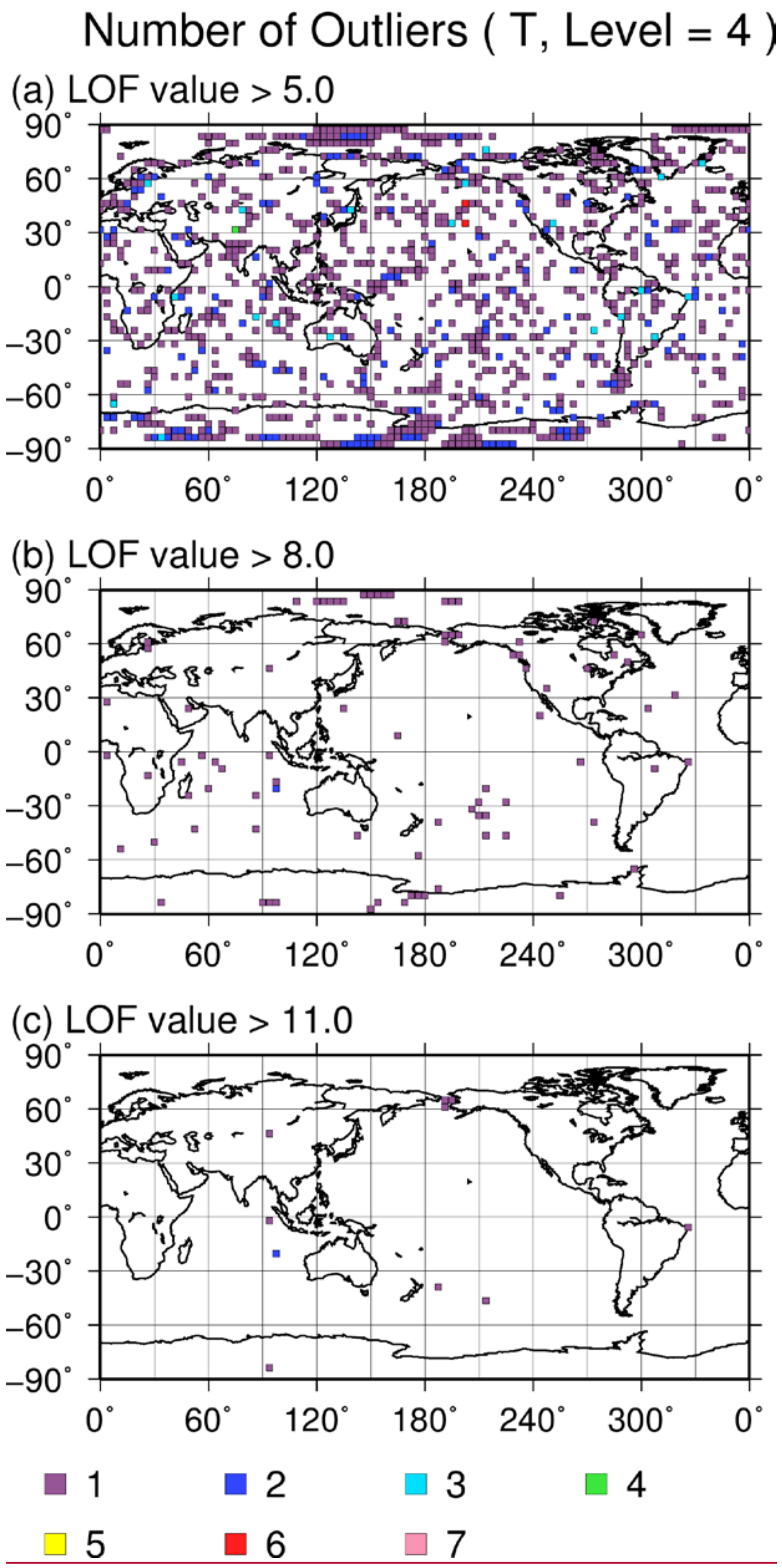


Figure 6: Spatial distributions of the number of outliers for background temperature at the fourth model level (~500 hPa) at 0600 UTC 22 February for *LOF* thresholds of (a) 5.0, (b) 8.0, and (c) 11.0.

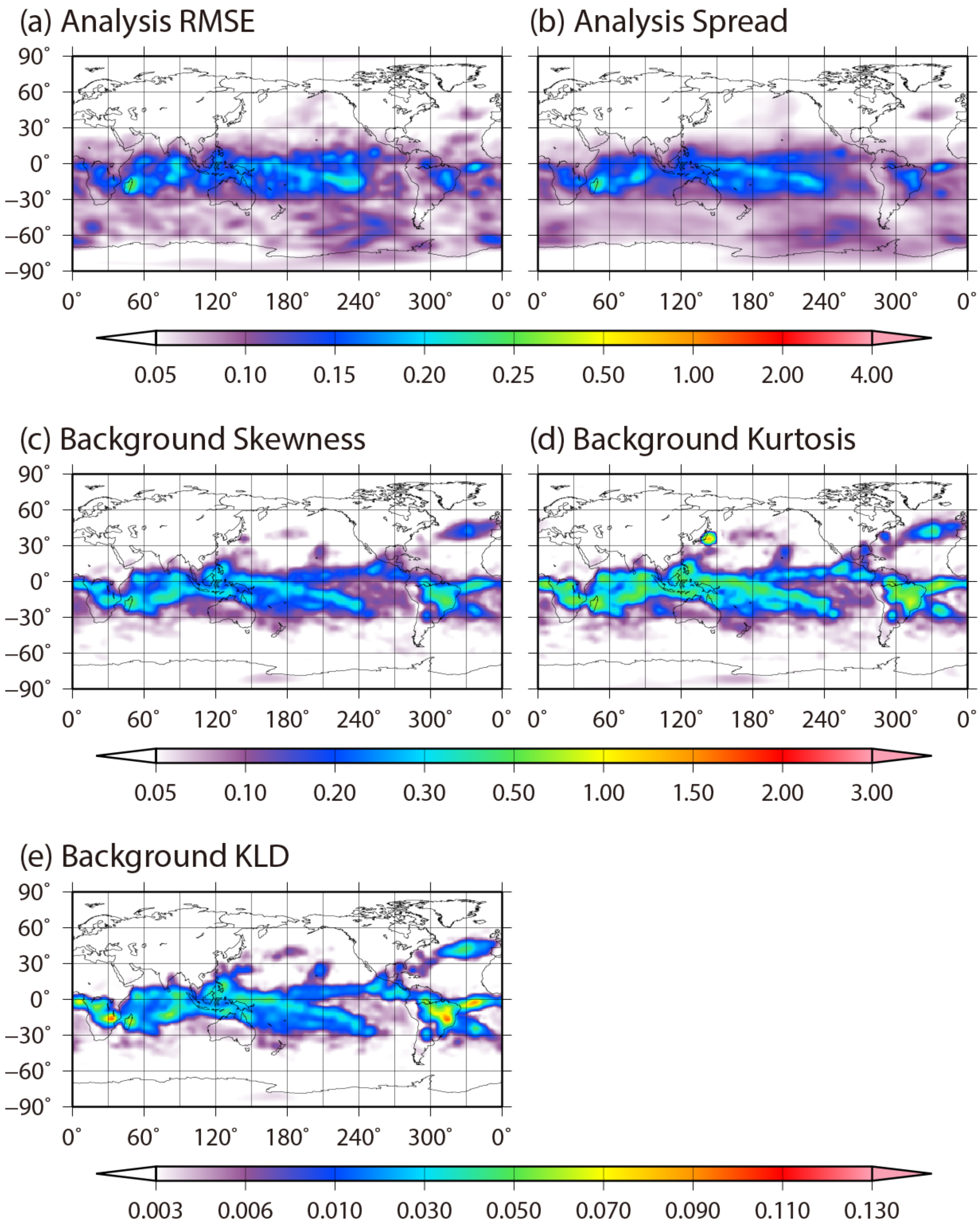


Figure 7: Spatial distributions of the time-mean (a) analysis RMSE, (b) analysis ensemble spread, (c) background absolute skewness, (d) background absolute kurtosis, and (e) background KL divergence for temperature at the fourth model level (~500 hPa) from 0000 UTC 25 January to 1800 UTC 1 March.

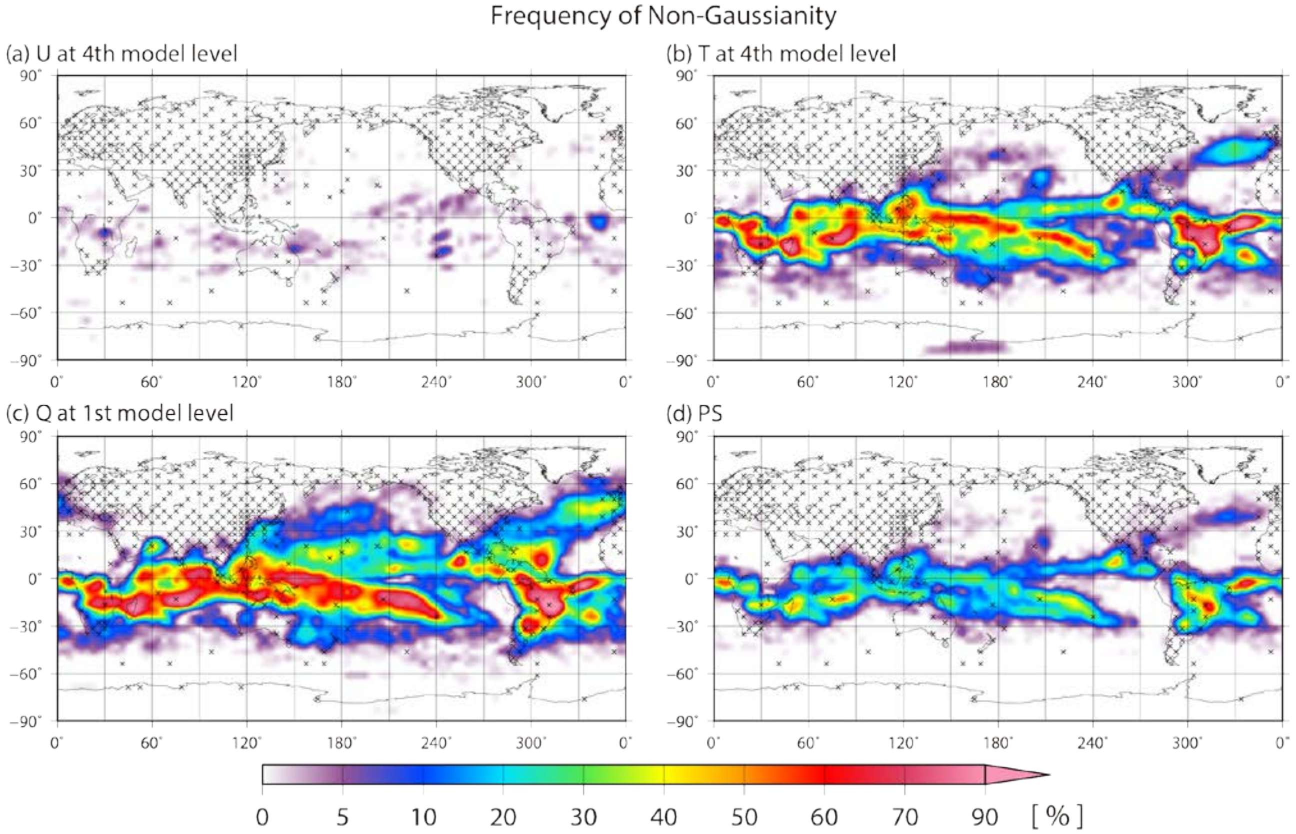


Figure 8: Spatial distributions of frequency of high KL divergence $D_{KL} > 0.01$ for (a) zonal wind at the fourth model level, (b) temperature at the fourth model level, (c) specific humidity at the lowest model level, and (d) surface pressure. The frequency is defined as a ratio of high KL divergence D_{KL} appearance from 0000 UTC 25 January to 1800 UTC 1 March.

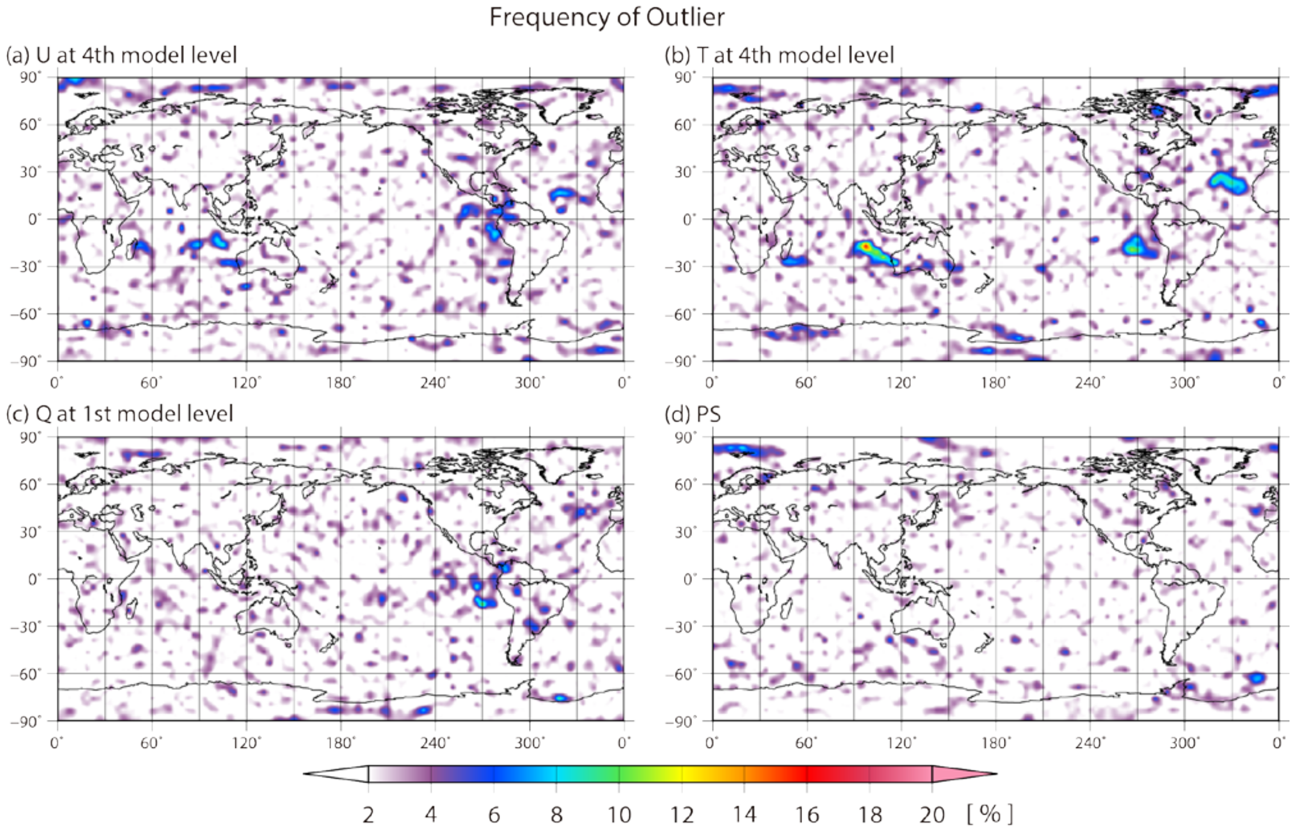


Figure 9: ~~As in~~Similar to Fig. ~~7~~8, but showing the frequency of high *LOF* ~~≥ 8, i.e., an outlier~~> 8 as outliers.

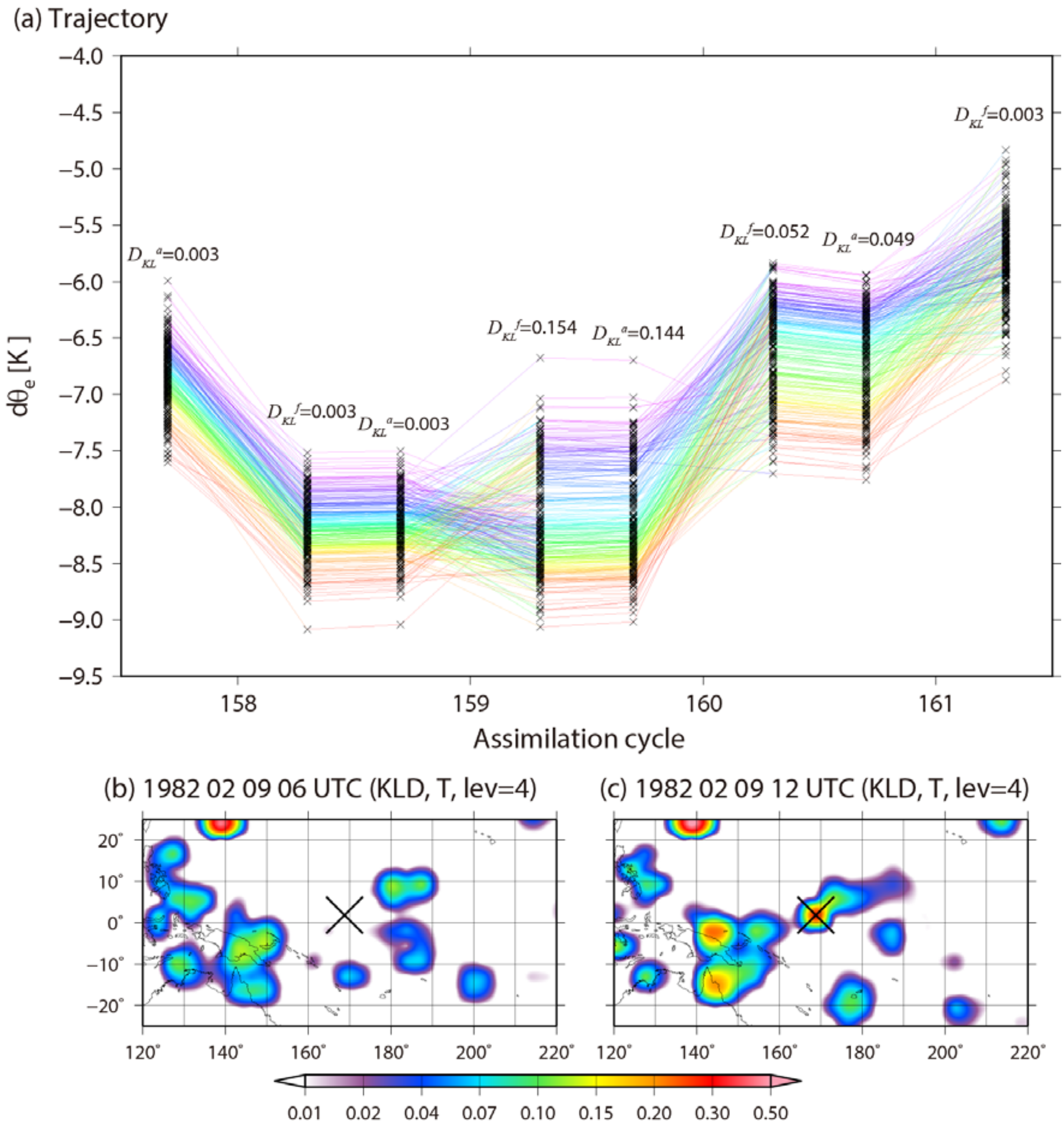


Figure 10: Lifecycle of non-Gaussianity at 1.856°N, 168.7°E. (a) Trajectories of 256 randomly chosen members from 10240 members for $d\theta_e$ ~~at 1.856°N, 168.7°E~~ from analysis at the 157th analysis cycle (0000 UTC 9 February) to forecast the 161st analysis cycle (0000 UTC 10 February). The colors show the order of $d\theta_e$ for every analysis. D_{KL} shows KL divergence for $d\theta_e$, and the superscripts *a* and *f* indicate analysis and forecast, respectively. (b, c) Spatial distributions of KL divergence for background temperature at the fourth model level (~500 hPa) at the 158th analysis cycle (0600 UTC 9 February) and the 159th analysis cycle (1200 UTC 9 February), respectively.

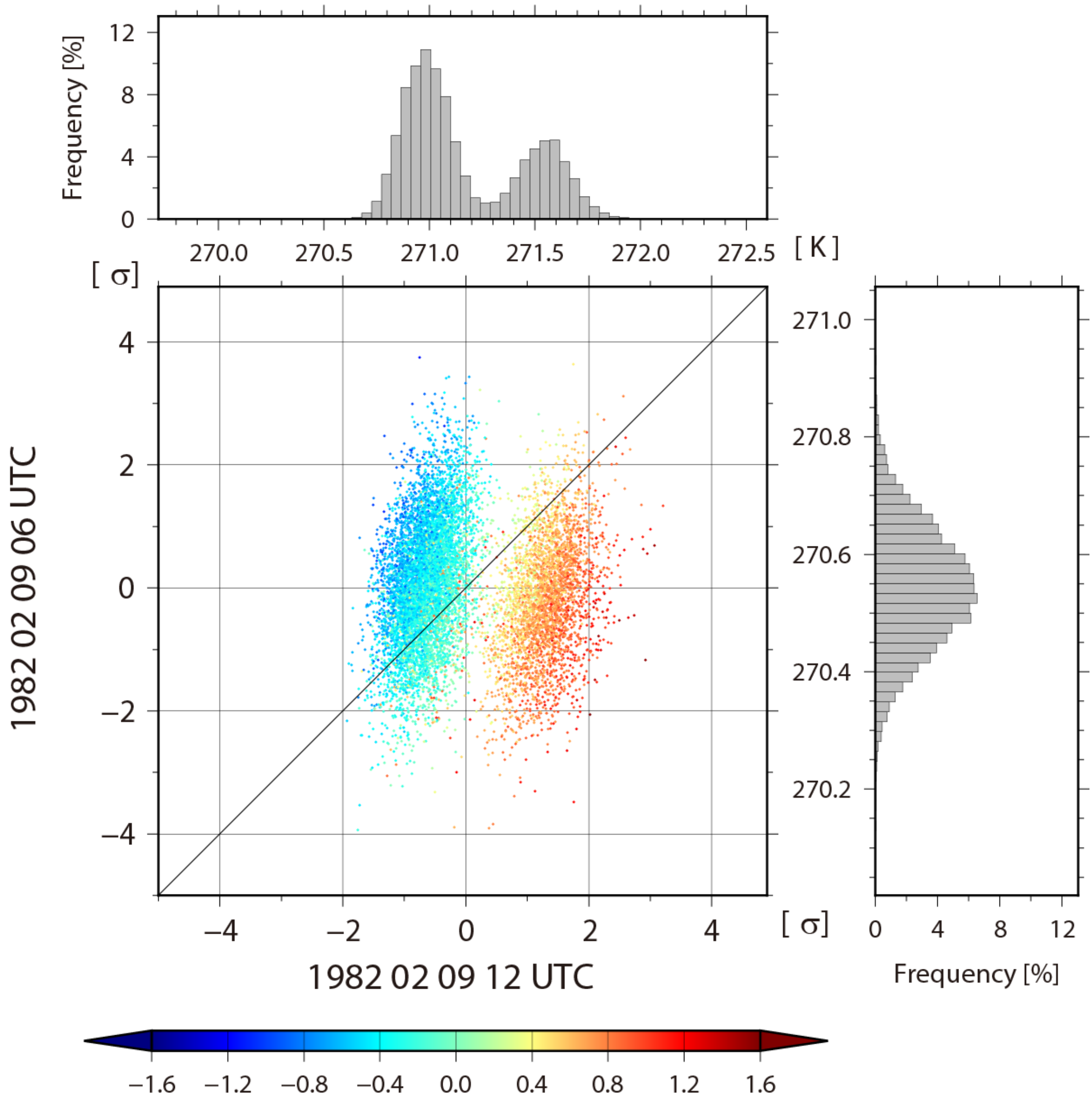


Figure 11: Scatter diagram of 0600 UTC versus 1200 UTC 9 February for the background temperature at the fourth model level (~500 hPa) at 1.856°N, 168.7°E. The colors show $d\theta'_e = (d\theta_{e\,1200\,UTC} - d\theta_{e\,0600\,UTC}) - (d\bar{\theta}_{e\,1200\,UTC} - d\bar{\theta}_{e\,0600\,UTC})$. The histograms on the ~~left~~right side and upper side show the background temperature at the same grid point.

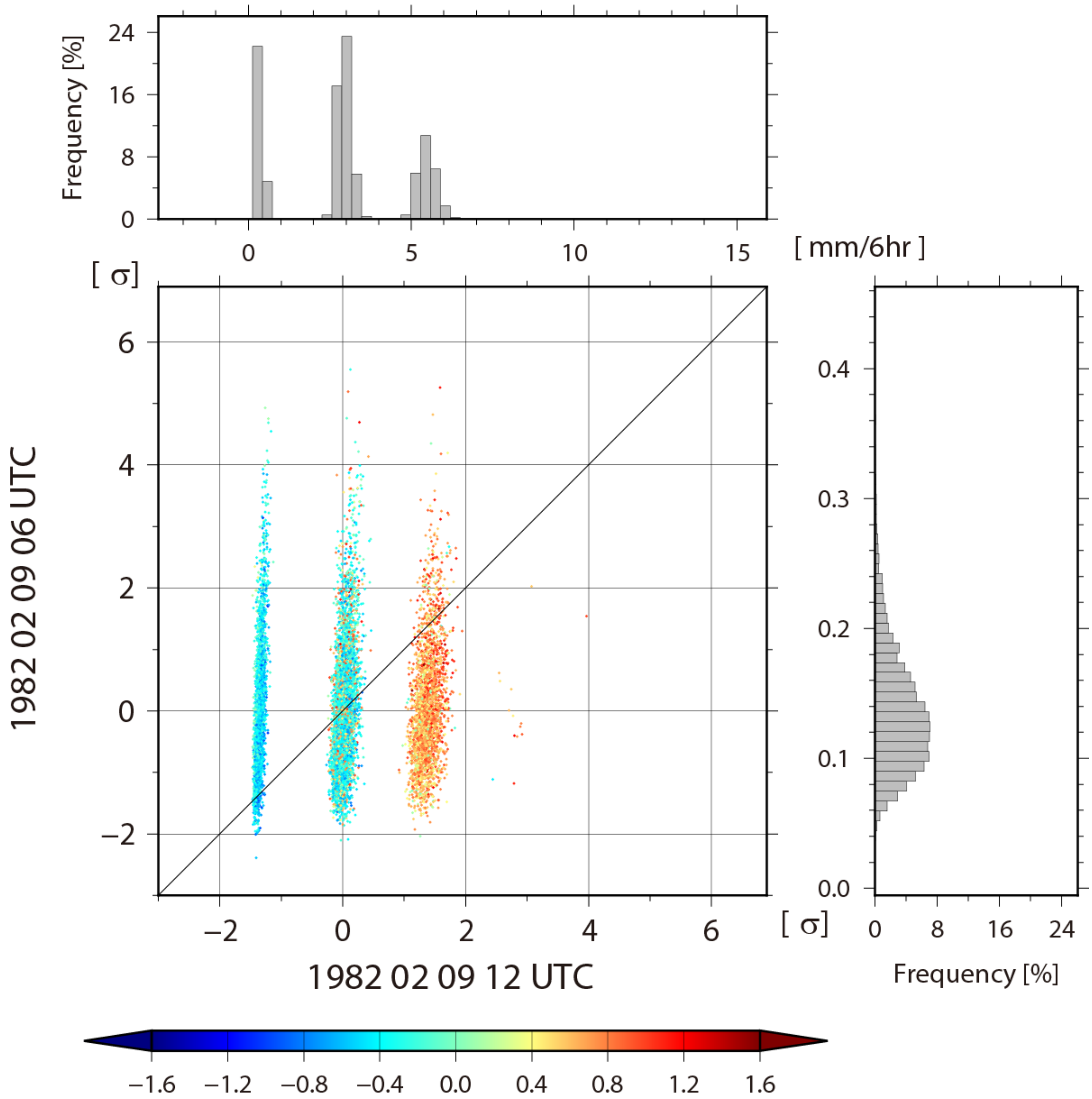


Figure 12: ~~As in~~Similar to Fig. ~~10~~11, but ~~background precipitation at~~for 0600 UTC versus ~~background temperature at~~ 1200 UTC 9 February for background precipitation.

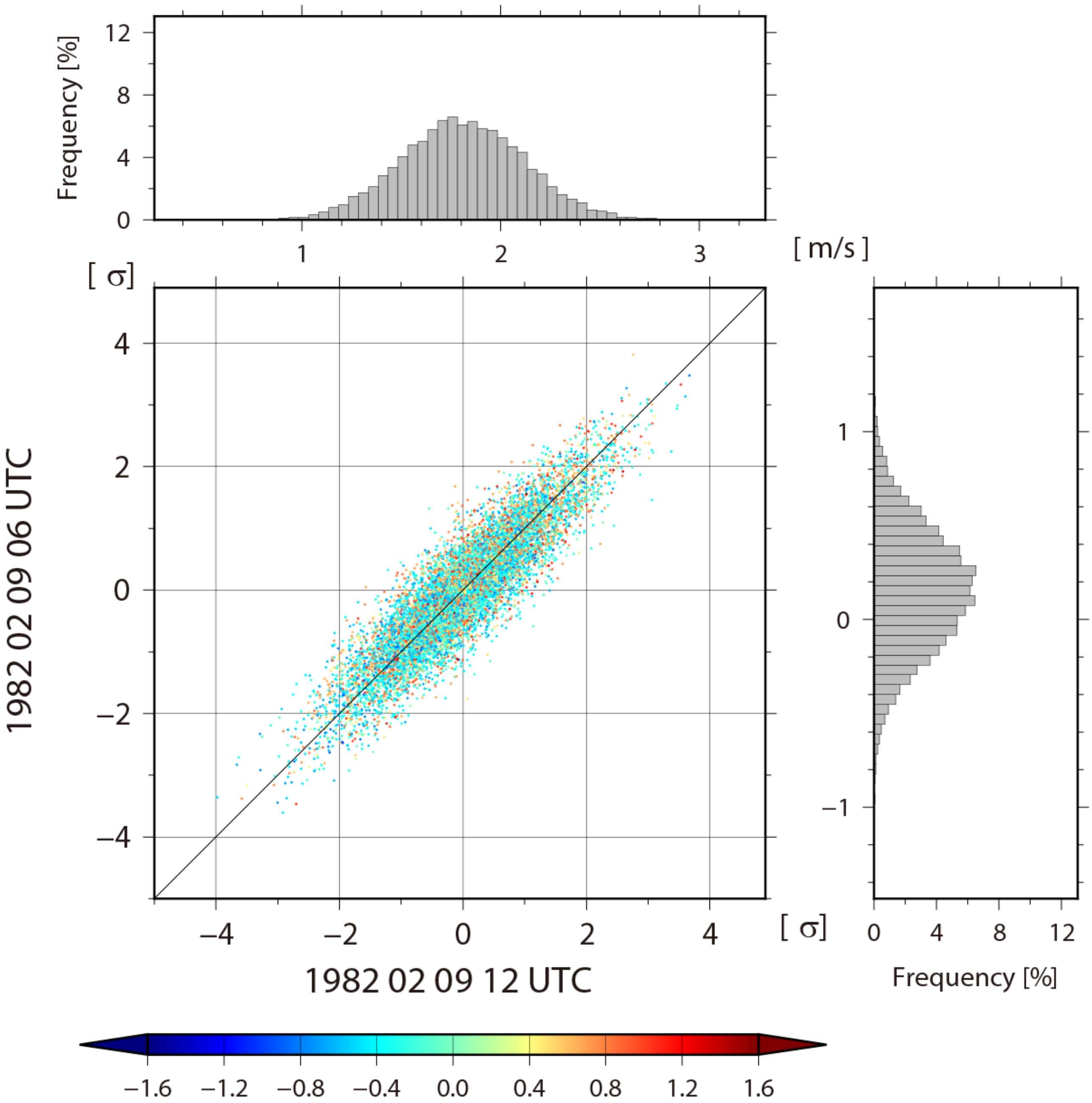


Figure 13: ~~As in~~Similar to Fig. ~~10~~11, but ~~background zonal wind at~~ for 0600 UTC versus ~~background temperature at~~ 1200 UTC 9 February~~.~~ for background zonal wind at the fourth model level (~500 hPa).

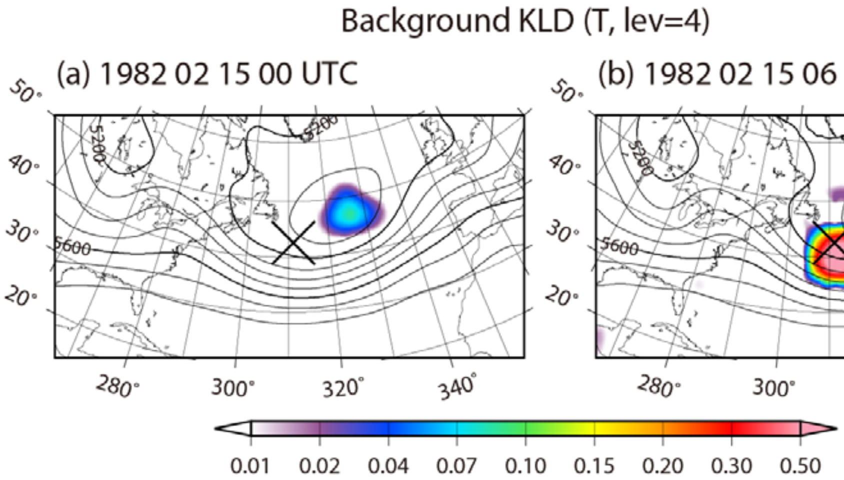


Figure 14: Spatial distributions of the KL divergence for background temperature at the fourth model level (~500 hPa) (a) at 0000 UTC 15 February and (b) at 0600 UTC 15 February. Contours show geopotential height of the ensemble mean at the 500 hPa level.

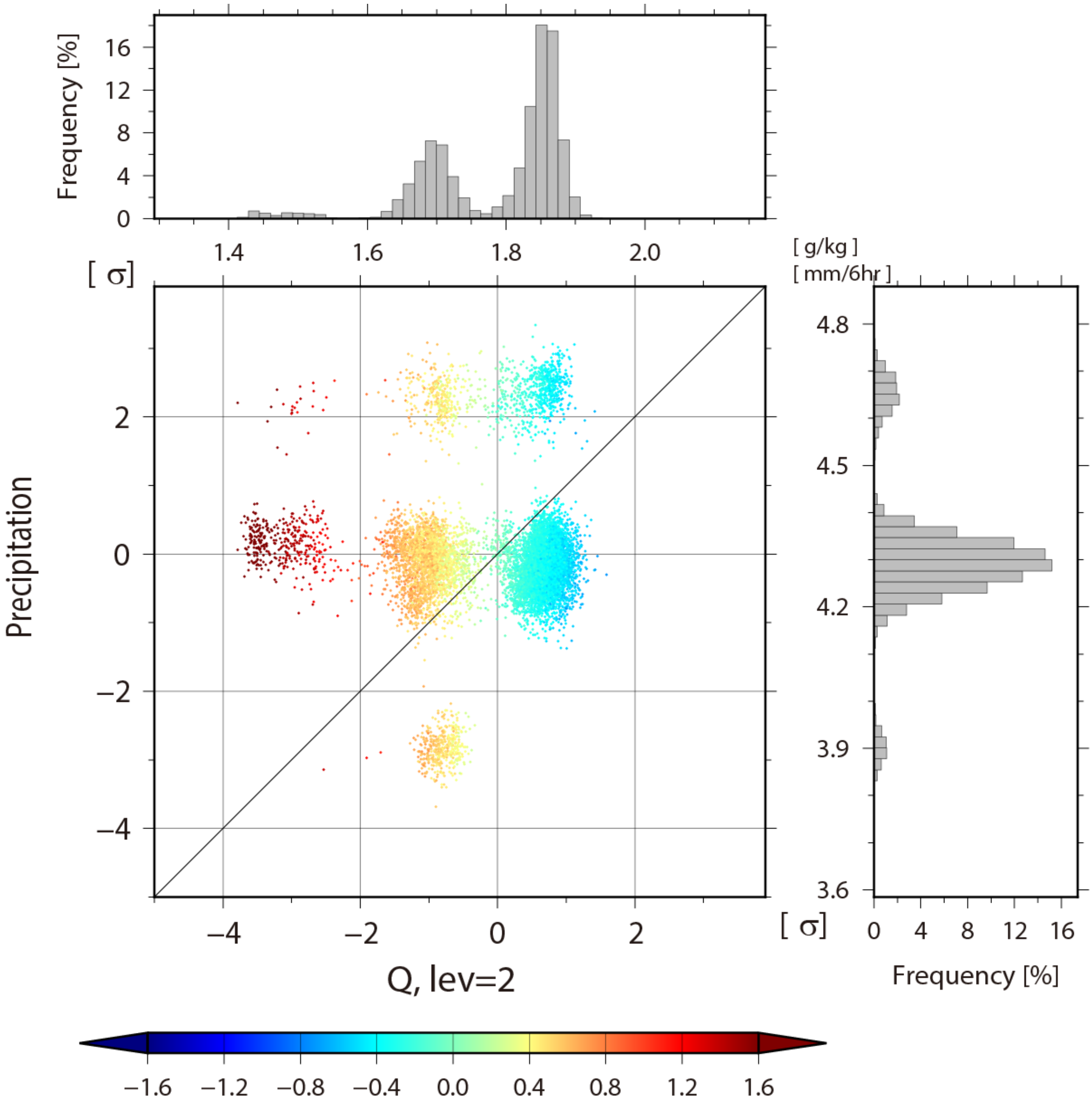


Figure 15: Scatter diagram of background specific humidity at the second model level (~850 hPa) versus background precipitation at 42.678°N, 48.75°W (311.25°E) at 0600 UTC 15 February. The colors show $d\theta_e' = (d\theta_{e\ 0600\ UTC} - d\theta_{e\ 0000\ UTC}) - (d\bar{\theta}_{e\ 0600\ UTC} - d\bar{\theta}_{e\ 0000\ UTC})$. The histograms on the right side and on top show background precipitation and temperature at the same grid point, respectively.

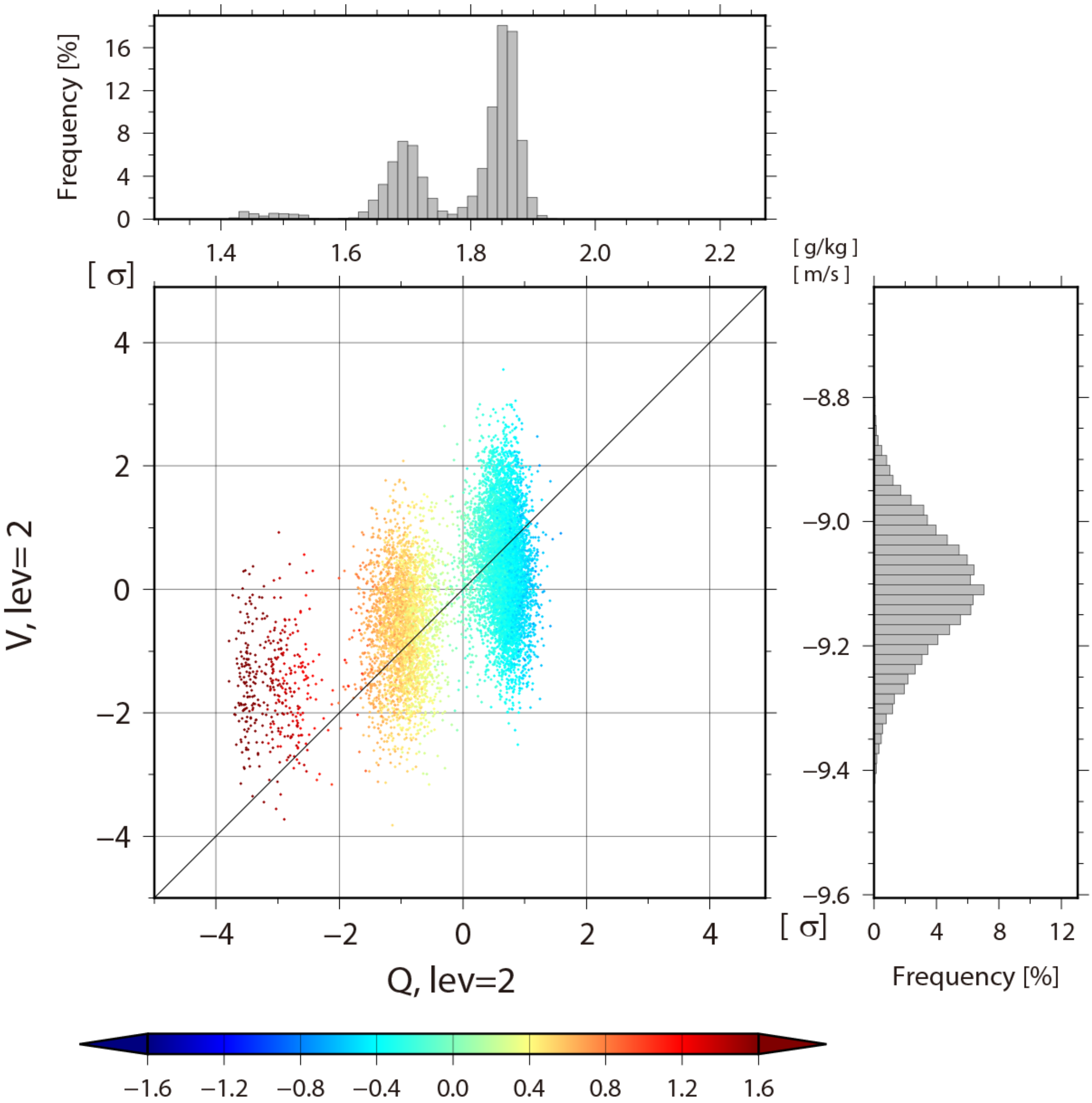
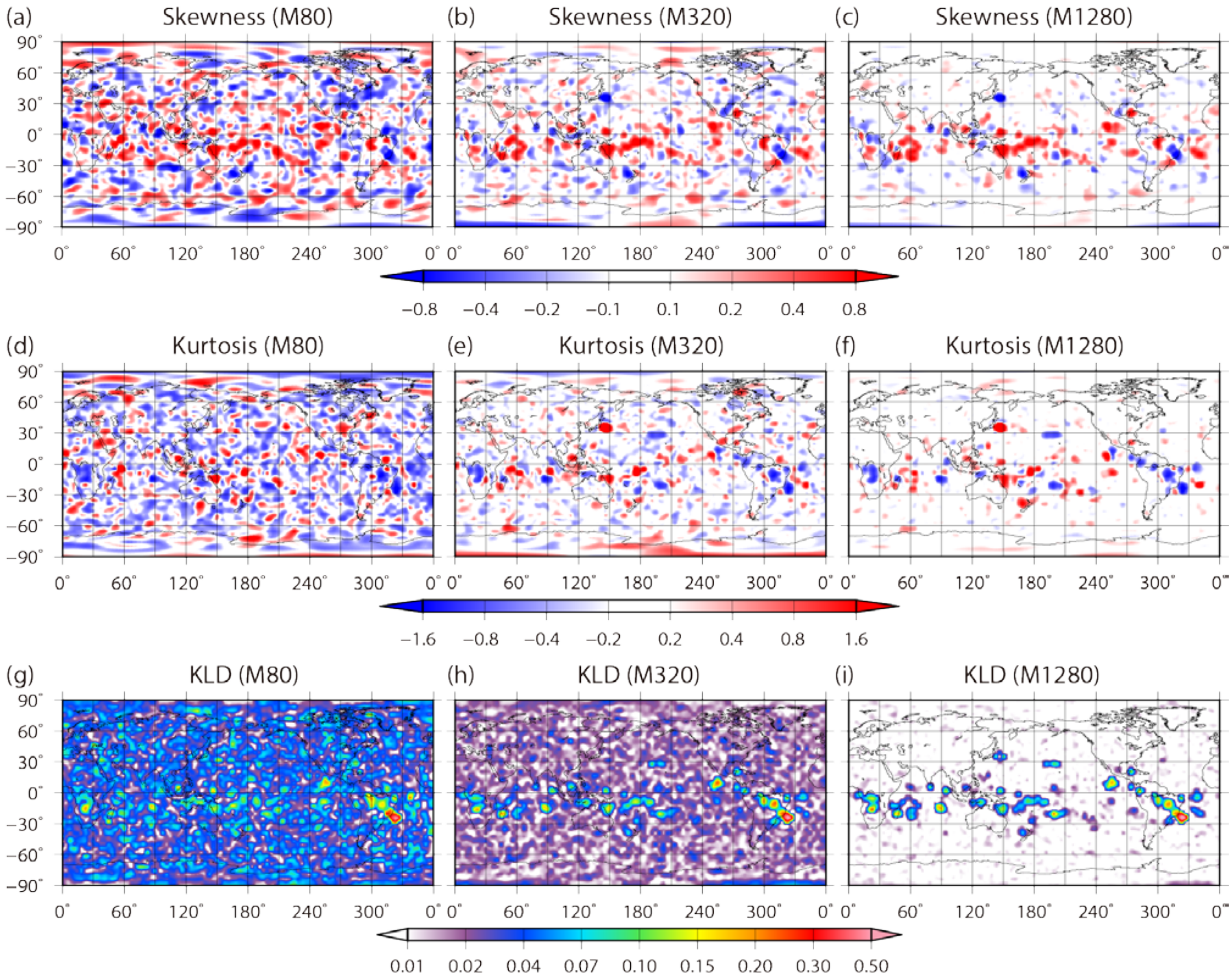
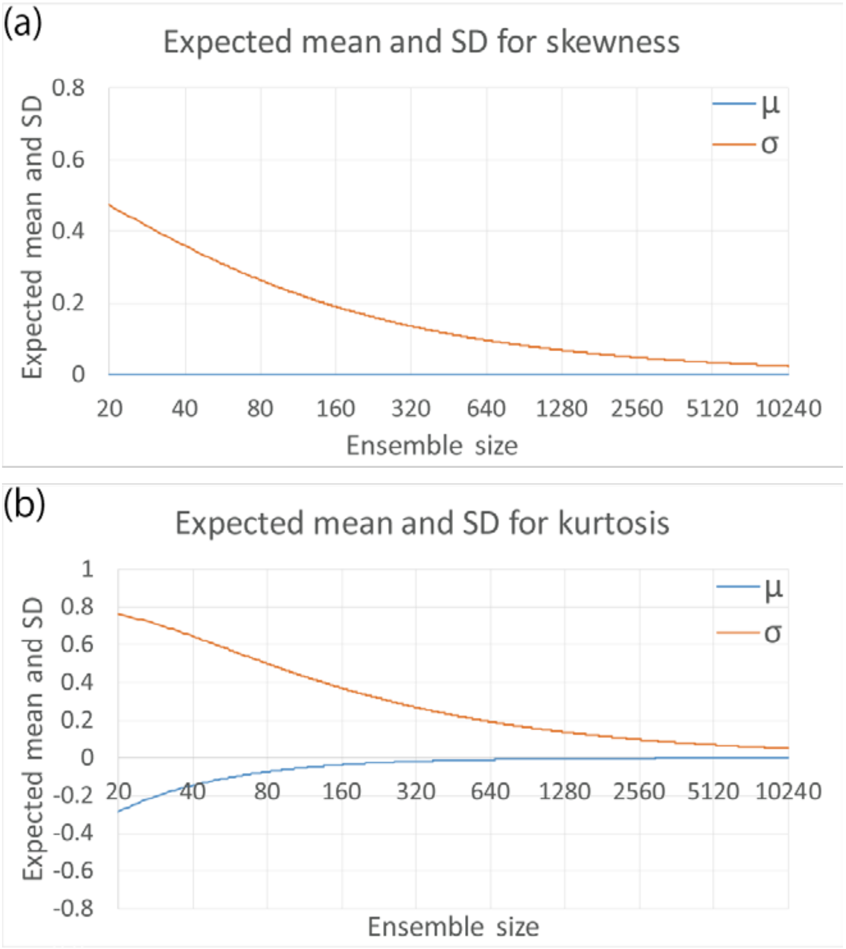


Figure 16: Similar to Fig. 14, but for background specific humidity versus meridional wind background at the second level (~850 hPa).



~~Figure 14~~Figure 17: Spatial distributions of (a-c) skewness, (d-f) kurtosis, and (g-i) KL divergence for temperature at the fourth model level (~500 hPa) at 0600 UTC 22 February. The left, center, and right columns show 80, 320, and 1280 subsamples from 10240 members, respectively.



~~Figure 15: Expected means and standard deviations of (a) skewness $\beta_1^{1/2}$ and (b) kurtosis β_2 with increasing ensemble size up to 10240. Blue and orange curves show the mean and standard deviation, respectively.~~

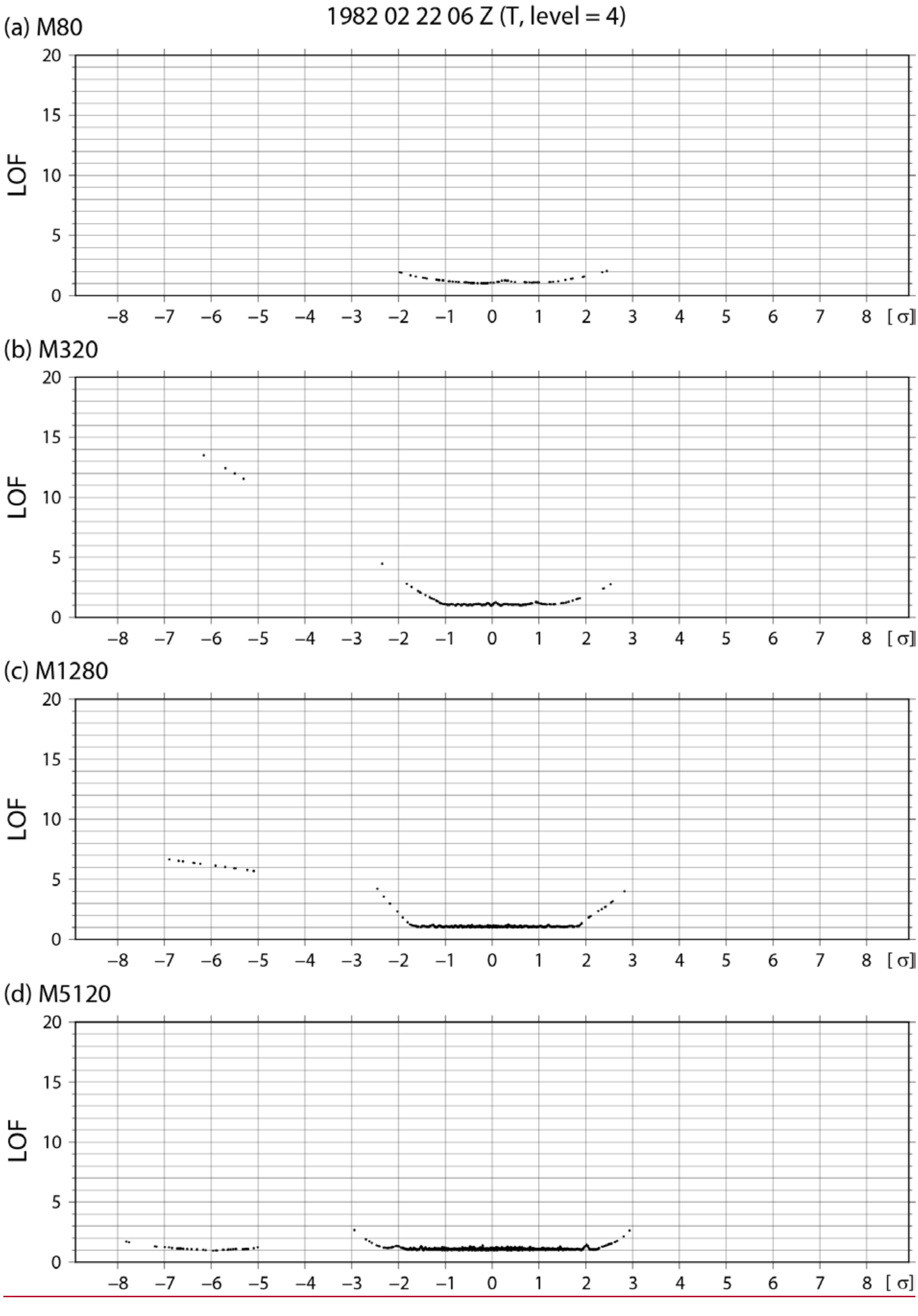

1982 02 22 06 Z (T, level = 4)
(a) M80
(b) M320
(c) M1280
(d) M5120
LOF
0
5
10
15
20
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
7
8
[σ]

Figure ~~16: As in~~ 18: Similar to Fig. 5b, but for the ensemble ~~size~~sizes (a) 80, (b) 320, ~~and~~ (c) 1280, and (d) 5120.

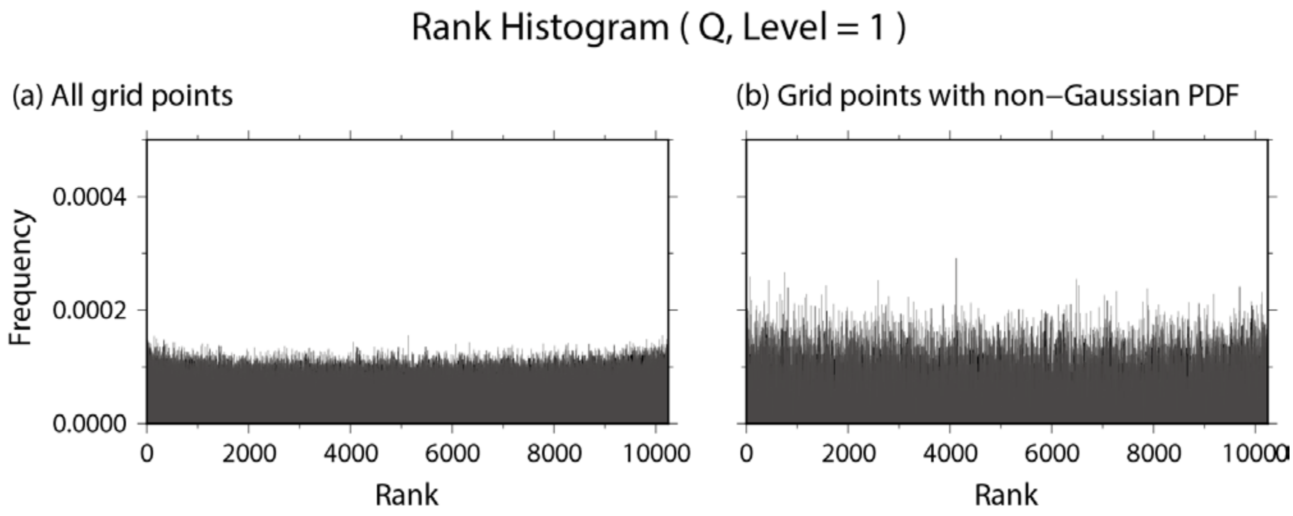


Figure 19: Rank histograms verified against truth for background specific humidity at the lowest model level (~925 hPa) at (a) all grid points and (b) the grid points with non-Gaussian PDF from 0000 UTC 25 January to 1800 UTC 1 March.

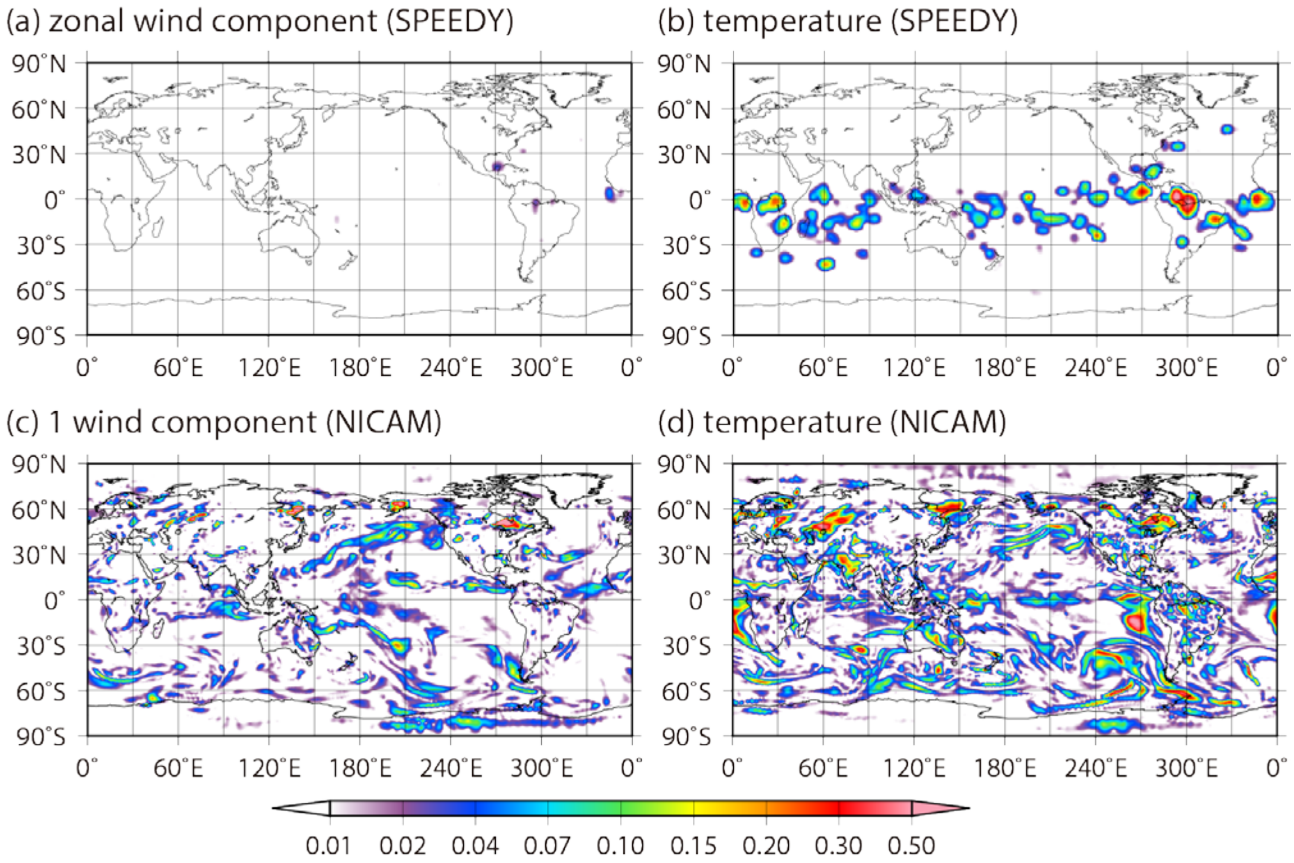


Figure ~~17~~20: Spatial distributions of background KL divergence for SPEEDY model and NICAM. Upper panels show (a) zonal wind and (b) temperature at the second model level (~850 hPa) for the SPEEDY model at 0000 UTC 1 March. Bottom panels show (c) one of three horizontal wind components and (d) temperature at the fifth model level (~850 hPa) for NICAM at 0000 UTC 8 November 2011.

Table. 1: CRPS and its three components (reliability, resolution and uncertainty) for background specific humidity at the lowest model level (~925 hPa) from 0000 UTC 25 January to 1800 UTC 1 March.

	CRPS [g kg^{-1}]	Reli [g kg^{-1}]	Resol [g kg^{-1}]	*U* [g kg^{-1}]
All grid points	0.0214	0.0000101	0.525	0.547
Grid points with non-Gaussian PDF	0.0475	0.0000244	0.030	0.077