Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018


Nonlinear Processes in Geophysics
Discussions

# 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 experiments with up to 10240 ensemble members using an intermediate atmospheric general circulation model. While the previous study focused on the 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 where highly nonlinear convective processes appear frequently. Therefore, we further investigate the lifecycle of multimodal PDFs, and show that the multimodal PDFs are generated by the on-off switch of convective parameterization and disappear naturally. Sensitivity to the ensemble size suggests that approximately 1000 ensemble members be necessary to capture the detailed structures of the non-Gaussian PDF.

## 1 Introduction

Data assimilation is a statistical approach to find the optimal initial state in 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) 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). 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 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. 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 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 how many ensemble members are necessary to capture non-Gaussian PDF. 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 $\beta_1^{1/2}$ and kurtosis $\beta_2$ are well-known parametric properties of a non-Gaussian PDF, and are defined as follows:

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

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

where $x_i$ and $\bar{x}$ denote the *i*th ensemble member and *N*-member ensemble mean, respectively; $\sigma$ denotes the sample standard deviation, i.e., $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$; 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 $\beta_2$ corresponds to the PDF with a more pointed (rounded) peak and longer (shorter) tails on both sides. When the PDF is Gaussian, skewness $\beta_1^{1/2}$ and kurtosis $\beta_2$ are both zero. 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 $\sigma$. The KL divergence 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 $\sigma$, respectively. $D_{KL}$ measures

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018

Nonlinear Processes in Geophysics Discussions
Open Access

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 $\beta_2$, and KL divergence $D_{KL}$. Here, the Scott's choice method (Scott 1979) is applied to decide the bin width for histograms. In this study, 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; Breunig et al. 2000).

In the SD method, the ensemble members beyond a prescribed threshold in the unit of SD are defined as outliers. If we have 10240 samples from the Gaussian PDF, 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. Using $\pm 3\sigma$ and $\pm 4\sigma$ thresholds, the outliers appear too frequently because 100% and 65% of all grid points statistically have at least one outlier under the Gaussian PDF. Therefore, we set the threshold as $\pm 5\sigma$ in this study.

The LOF method is based on local density, and not distance as in the SD method. For a given 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:

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

The cardinality of $N_k(p)$, or $|N_k(p)|$, is greater than or equal to the number of objects except for the

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018


Nonlinear Processes in Geophysics
Open Access
Discussions

object $p$ 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 $k\text{-}distance(o)$ 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 object with *LOF* much larger than 1 may be categorized as an outlier, it is not clear how to determine the threshold for outliers because the threshold also depends on the dataset. In Section 4, we describe the threshold in detail. $k$ is a control parameter for the LOF method, and depends on the dataset (Breunig et al. 2000). Markus et al. (2000) suggested that choosing $k$ from 10 to 20 work well for most of the datasets; we chose $k = 20$ in this study.

## 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

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018




K supercomputer. The LETKF is computed as follows. Let $\mathbf{X}$ ($\delta\mathbf{X}^f$) denote an $n \times m$ matrix, the columns of which are composed of $m$ ensemble members (ensemble perturbations) with the system dimension $n$. The analysis ensemble $\mathbf{X}^a$ is written as:

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

[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 $\widetilde{\mathbf{P}}^a$ in the ensemble space is given as

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

[cf. Eqs. (3) and (9) of Miyoshi and Yamane 2007]. Here, $\rho$ denotes the covariance inflation factor. As $\widetilde{\mathbf{P}}^a$ 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. 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 $\beta_2$, 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 $\beta_2$, 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 $\beta_2$, 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 Gaussian, skewness $\beta_1^{1/2}$ and kurtosis

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018



$\beta_2$ are extremely large. Moreover, the PDF does not fit 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 $-5\sigma$ 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 $-5\sigma$, 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 $\beta_2$, 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. 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 $LOF \geq 8$, 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 frequency of high KL divergence $D_{KL}$ for temperature corresponds to the time mean RMSE and $D_{KL}$ (Figs. 7 a, e, and 8 b), and the pattern correlation 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 appears for 34 days out of 36 days. In contrast, the non-Gaussianity 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.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018



To investigate how the non-Gaussianity 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). That is, Fig. 10 shows the lifecycle of the non-Gaussianity. As the vertical axis, we introduce the convective instability $d\theta_e$, which is defined as a difference between equivalent potential temperature $\theta_e$ at the fourth model level (~500 hPa) and $\theta_e$ at the second model level (~850 hPa). Negative (Positive) $d\theta_e$ indicates a convectively unstable (stable) atmosphere. The non-Gaussianity appears in the background at the 159th cycle (1200 UTC 9 February), and the model forecast generates the obvious non-Gaussianity. The members of the upper side cluster at the 159th cycle generally become unstable in the forecast step and their instability is mitigated in the model. In contrast, most other members show enhanced instability. Finally, the non-Gaussianity 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 KL divergence $D_{KL}$ becomes large from 0.003 to 0.299 for 6 h. The dot colors show $d\theta_e' = (d\theta_{e\,1200\,UTC} - d\theta_{e\,0600\,UTC}) - (d\bar{\theta}_{e\,1200\,UTC} - d\bar{\theta}_{e\,0600\,UTC})$, 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

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018


right (left) side mode. The more outside members at 1200 UTC become much more stable (redder) or unstable (bluer), respectively. 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 the zonal wind. We could not find any relationships between the atmospheric instability and zonal wind. Therefore, the non-Gaussianity genesis is deeply related to precipitation process, which is driven by convective instability through cumulus parameterization. As a result, the precipitation process mitigates the instability, which raises the temperature and reduces the humidity. 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 shows the spatial distributions of the skewness $\beta_1^{1/2}$, kurtosis $\beta_2$, 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 $\beta_2$, and KL divergence $D_{KL}$ with 80 members contain high levels of contaminating errors originating from sampling errors, and

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018




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. Thus, a sample size of about 1000 members is sufficient to discuss non-Gaussianity. 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 $\beta_2$ under the Gaussian distribution. The expected mean and standard deviation of distribution of skewness $\beta_1^{1/2}$ are written as

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

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

and the expected mean and standard deviation of distribution kurtosis $\beta_2$ are given by

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

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

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 $\beta_2$ 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 $\beta_2$ is negative and convergence is relatively slow. Indeed, the distribution of kurtosis $\beta_2$ with 80 members is mostly covered by negative values (Fig. 14d). The outliers also depend on the sample size. Figure 16 shows LOF with 80, 320, and 1280 members for temperature at the fourth model level, 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 $LOF \geq 8$ are divided into outliers. Moreover, with an ensemble size of 1280, 11 members construct a small cluster, 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).

## 5 Summary and discussions

Gaussian filters, such as EnKFs, cannot treat non-Gaussian PDF properly. That is, the Gaussian filters cannot produce accurate analysis when significant non-Gaussianity exists. Therefore, this study investigated the non-Gaussianity of the atmosphere and its behavior using a SPEEDY-LETKF system with 10240 members. The non-Gaussian PDF appears frequently in areas where the RMSE and ensemble spread are large. Moreover, an ensemble size of about 1000 is large enough to resolve the non-Gaussian PDF and to mitigate the sampling error.

The non-Gaussian PDF appears frequently in the tropics and storm tracks over the Pacific and Atlantic Oceans, particularly for temperature and specific humidity, but not winds. The genesis of non-Gaussianity is explained by the convective instability. These results suggest that the non-Gaussianity 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.

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.

The non-Gaussianity is less frequent in the wind components not only on 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 temperature and specific humidity fields to 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 (Miyoshi et al. 2015). Figure 17 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

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018


November 2011. With NICAM, the non-Gaussianity appears globally not only in the temperature field but also in the wind component. This result 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 number of outliers is basically one, but sometimes the number is more than one. 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 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.

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 generated through physical processes has some physical significance and should not be divided into outliers.

The non-Gaussian measures are sensitive to the ensemble size (see Figs. 14, 16). 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 are grateful to the members of the Data Assimilation Research Team, RIKEN R-CCS for fruitful discussions. The SPEEDY-LETKF code is publicly available at 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.: 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.

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: 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.

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.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018




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.

Tiedtke, M: A comprehensive mass flux scheme for cumulus parameterization in large-scale

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018




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.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018



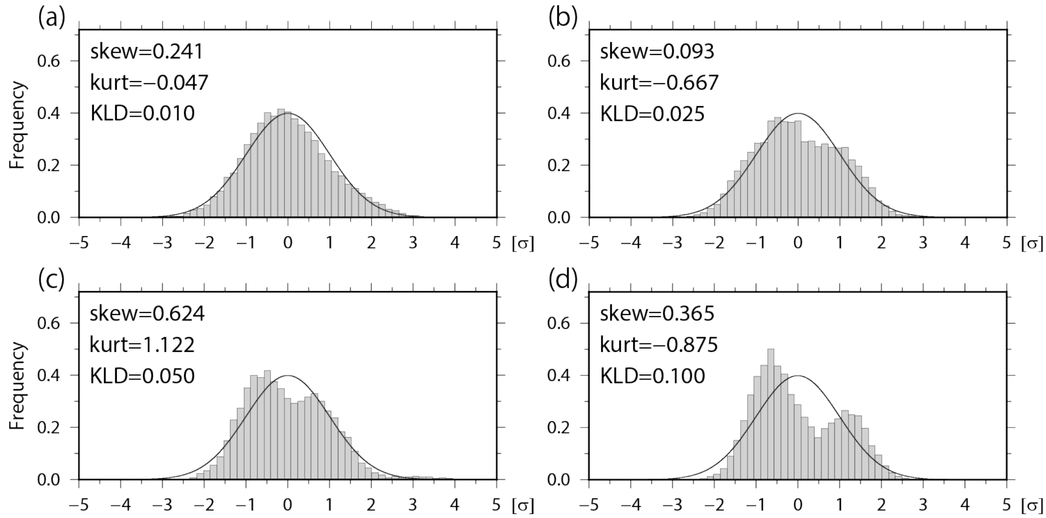



Figure 1: Ensemble-based histograms with 10240 ensemble members when the Kullback–Leibler (KL) divergence $D_{KL}$ = (a) 0.010, (b) 0.025, (c) 0.050, and (d) 0.100. Solid lines indicate fitted Gaussian functions. Skewness (skew) and kurtosis (kurt) are also shown in the figure.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018


Nonlinear Processes in Geophysics
Discussions
Open Access

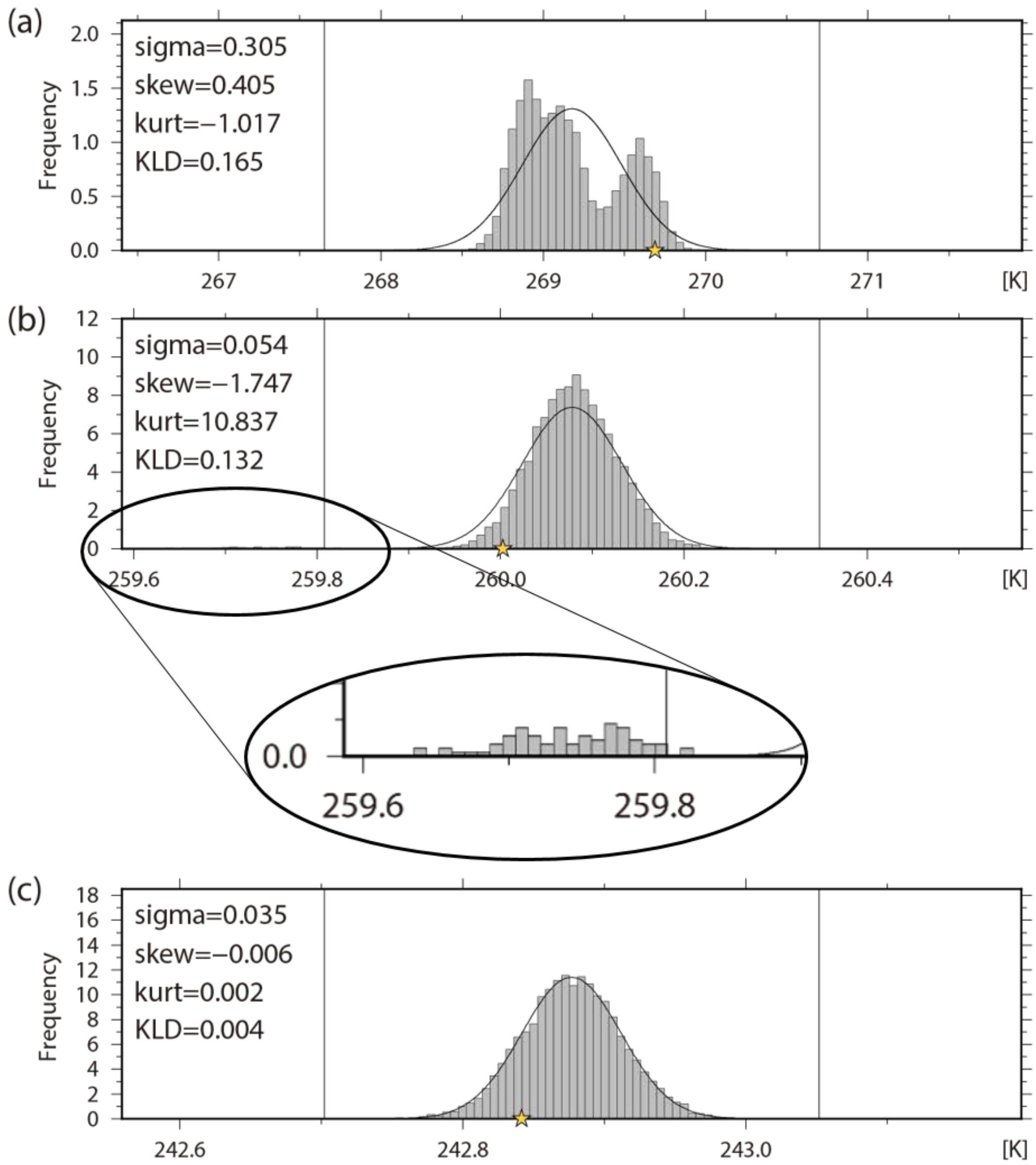



Figure 2: Histograms of temperature (K) at the fourth model level (~500 hPa) at (a) grid point A (16.7°S, 90.0°E), (b) grid point B (35.256°N, 146.25°E), and (c) grid point C (35.256°N, 112.5°W). The orange star shows the truth.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018


Nonlinear Processes in Geophysics
Open Access
Discussions

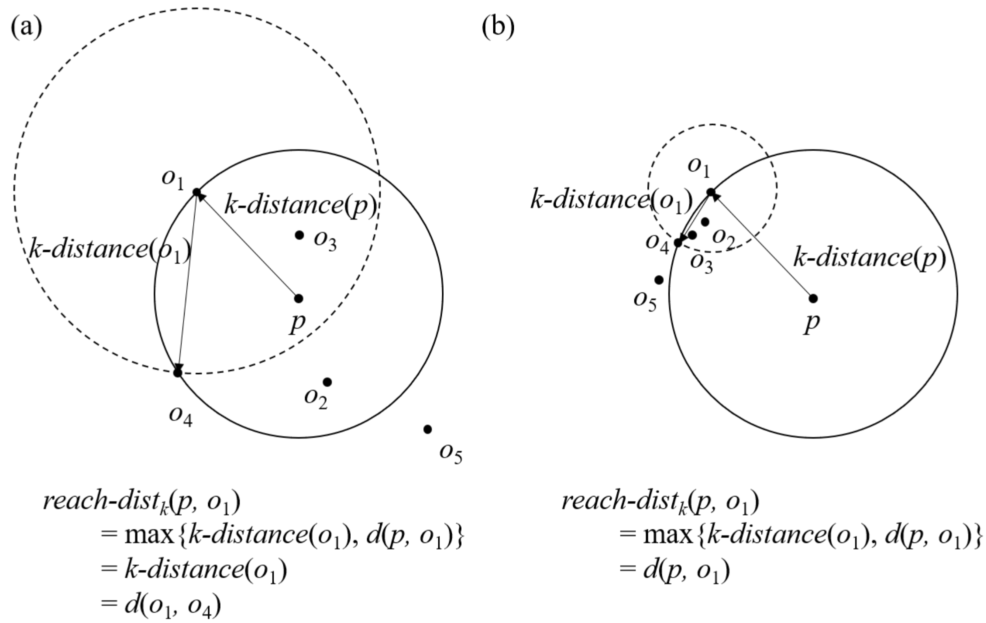



Figure 3: Schematic diagrams of *reach-dist*$_k$ $(p, o)$ with $k = 3$ for (a) uniformly distributed data and (b) data with an asymmetrical distribution.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018

Nonlinear Processes in Geophysics
Open Access
Discussions
EGU

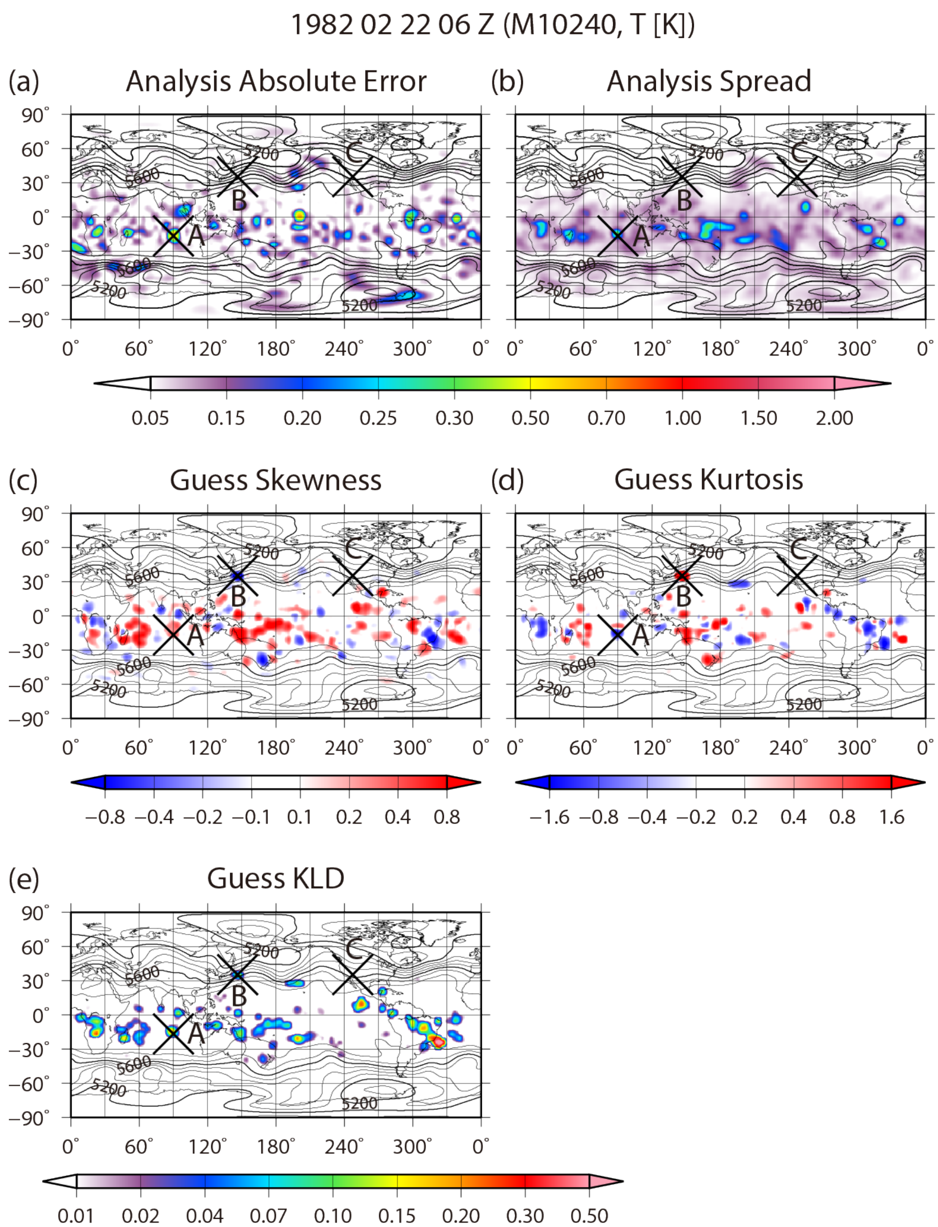



Figure 4: Spatial distributions of (a) analysis absolute error, (b) analysis ensemble spread, (c) background skewness, (d) background kurtosis, and (e) background KL divergence for temperature at the fourth model level (~500 hPa) on 0600 UTC 22 February. Contours indicate geopotential height of the ensemble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

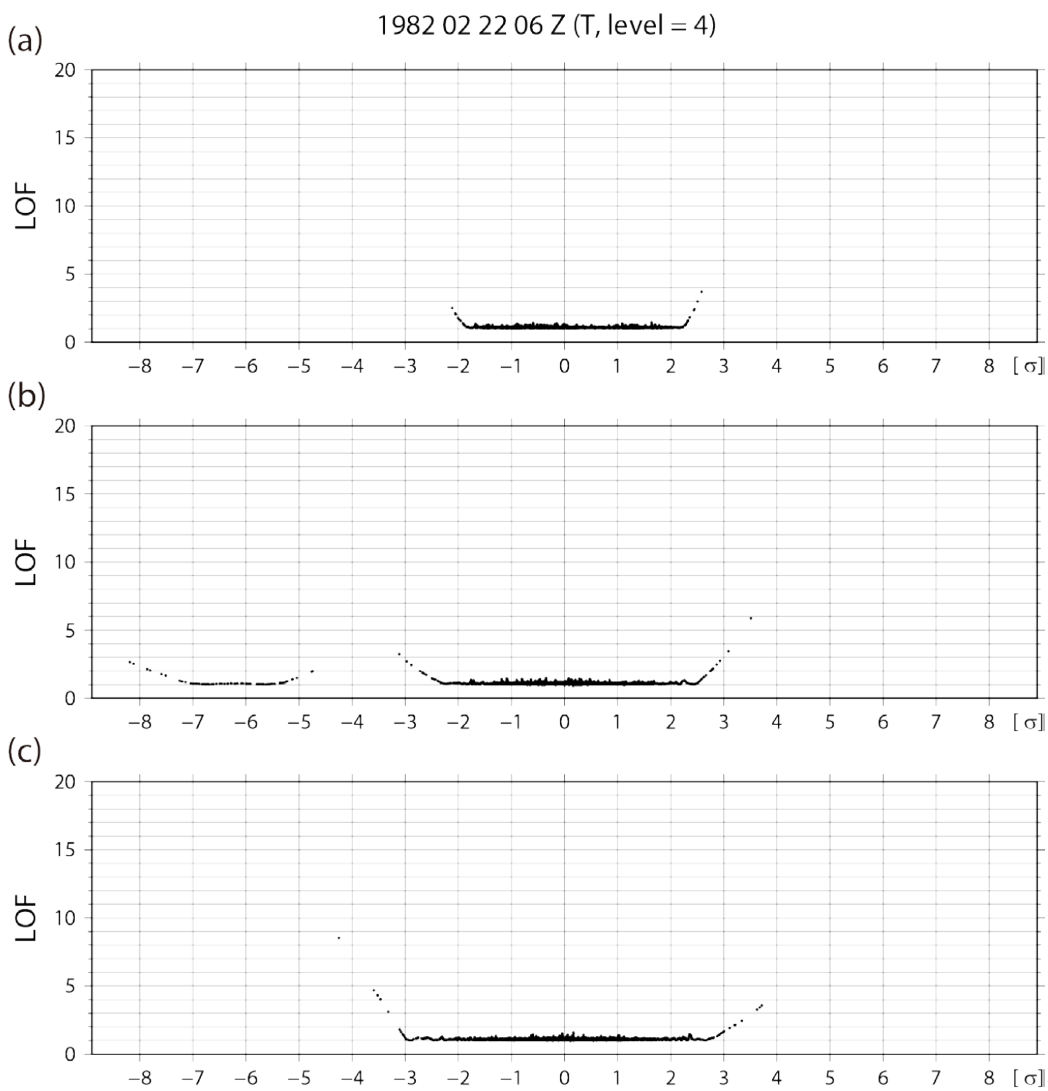



Figure 5: Scatter diagrams of the local outlier factor method (*LOF*) versus distance from the ensemble mean for all ensemble members for temperature at the fourth model level (~500 hPa) at (a) grid point A (16.7°S, 90.0°E), (b) grid point B (35.256°N, 146.25°E), and (c) grid point C (35.256°N, 112.5°W).

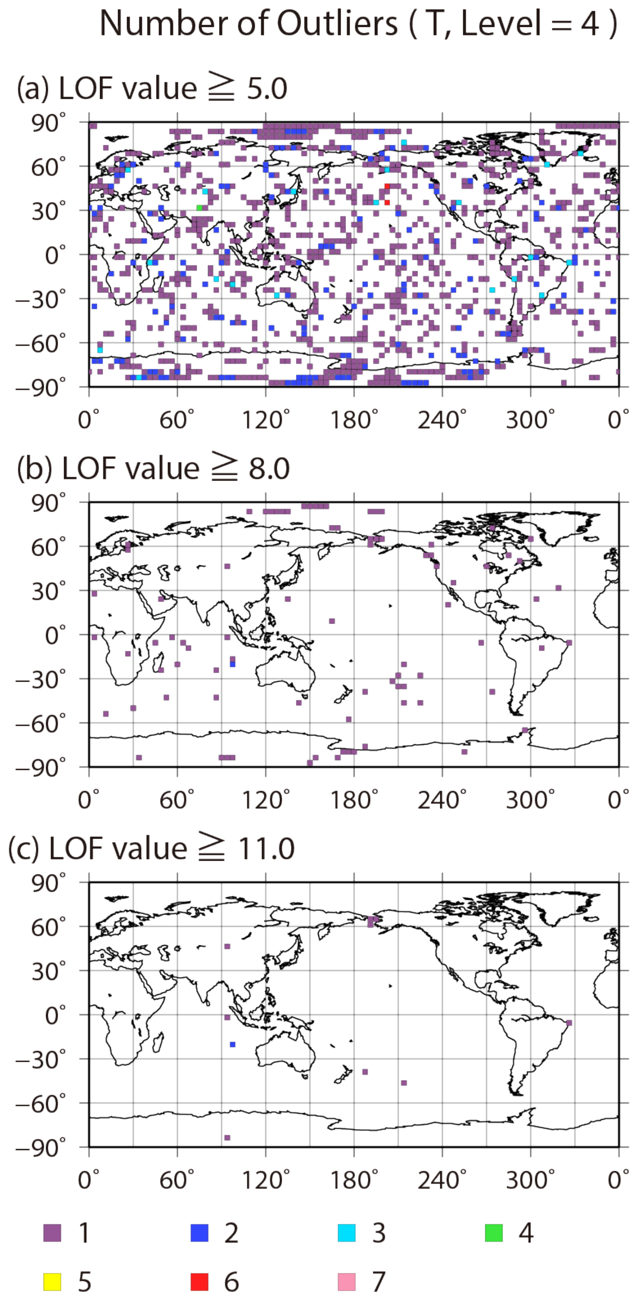



Figure 6: Spatial distributions of the number of outliers at 0600 UTC 22 February for *LOF* thresholds of (a) 5.0, (b) 8.0, and (c) 11.0.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018


Nonlinear Processes in Geophysics
Discussions

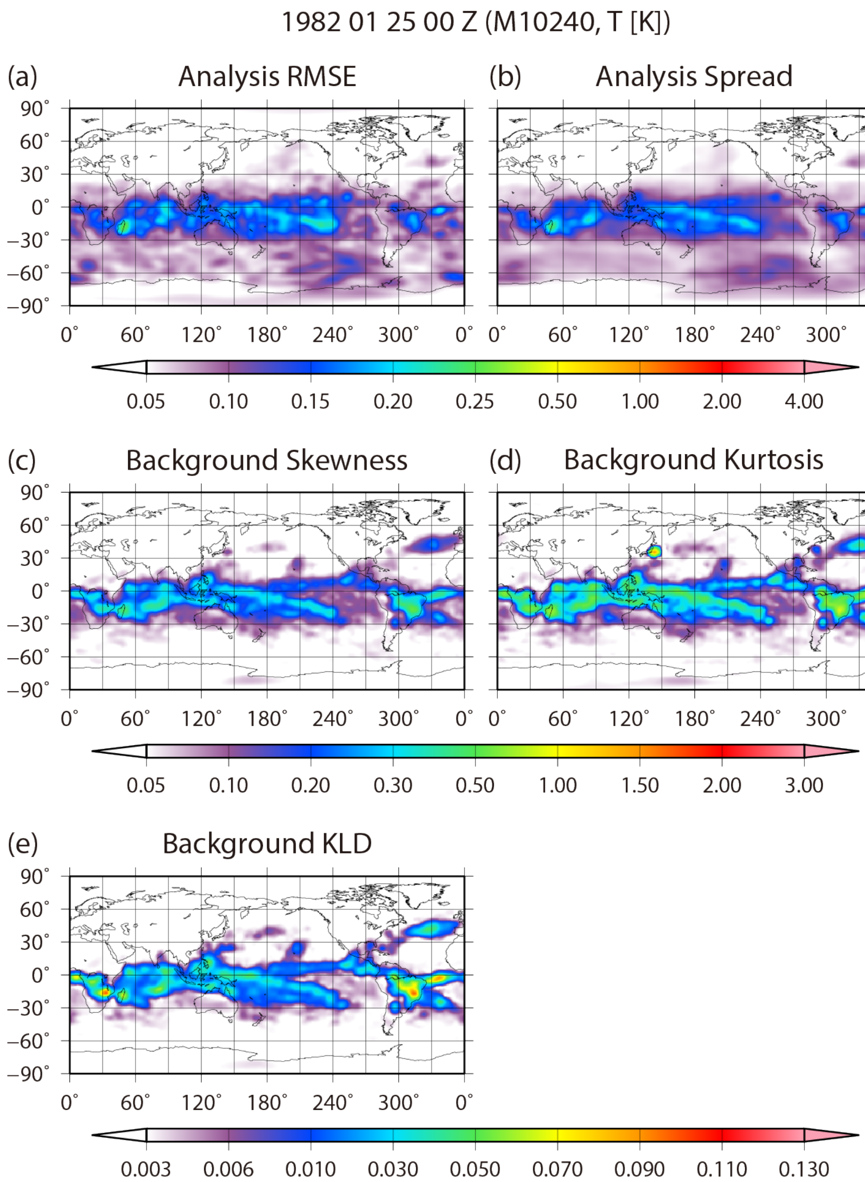



Figure 7: Spatial distributions of the time-mean (a) analysis RMSE, (b) analysis ensemble spread, (c) background absolute skewness, (d) background absolute kurtosis, and (e) background KL divergence for temperature at the fourth model level (~500 hPa) from 0000 UTC 25 January to 1800 UTC 1 March.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018


Nonlinear Processes in Geophysics
Discussions
Open Access

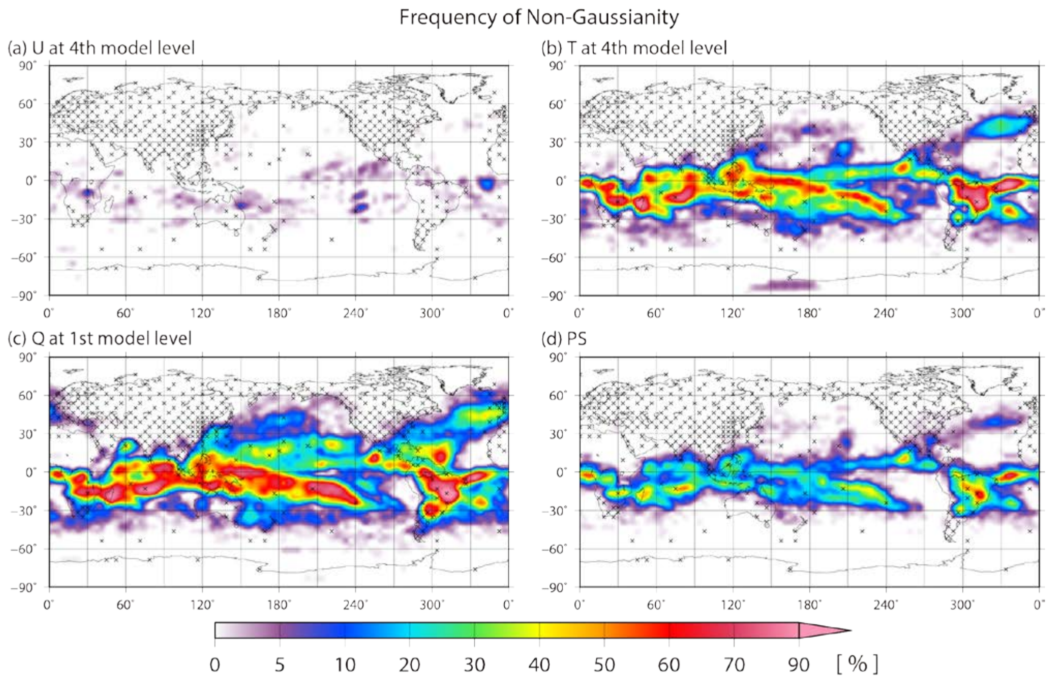



Figure 8: Spatial distributions of frequency of high KL divergence $D_{KL} > 0.01$ for (a) zonal wind at the fourth model level, (b) temperature at the fourth model level, (c) specific humidity at the lowest model level, and (d) surface pressure. The frequency is defined as a ratio of high KL divergence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

Nonlinear Processes in Geophysics Discussions
Open Access

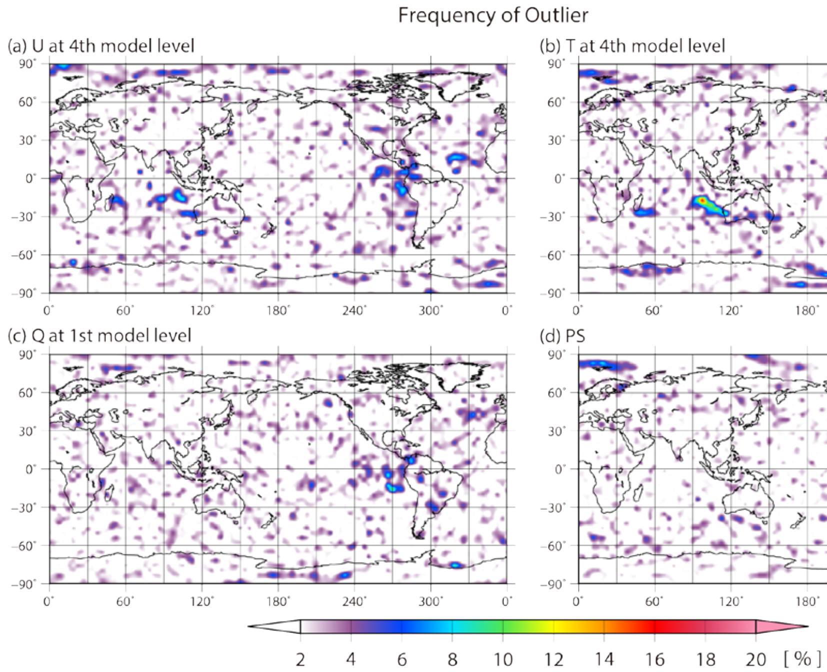



Figure 9: As in Fig. 7, but showing the frequency of high $LOF \geq 8$, i.e., an outlier.

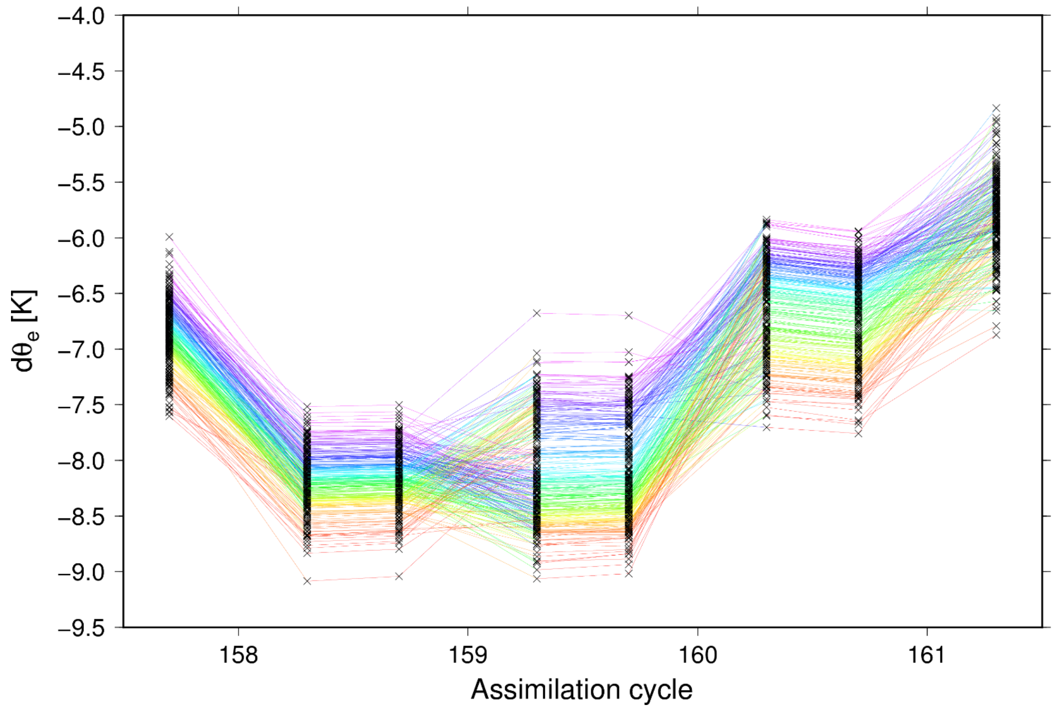



Figure 10: Trajectories of 256 randomly chosen members from 10240 members for $d\theta_e$ at 1.856°N, 168.7°E from analysis at the 157th analysis cycle (0000 UTC 9 February) to forecast the 161st analysis cycle (0000 UTC 10 February). The colors show the order of $d\theta_e$ for every analysis.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018



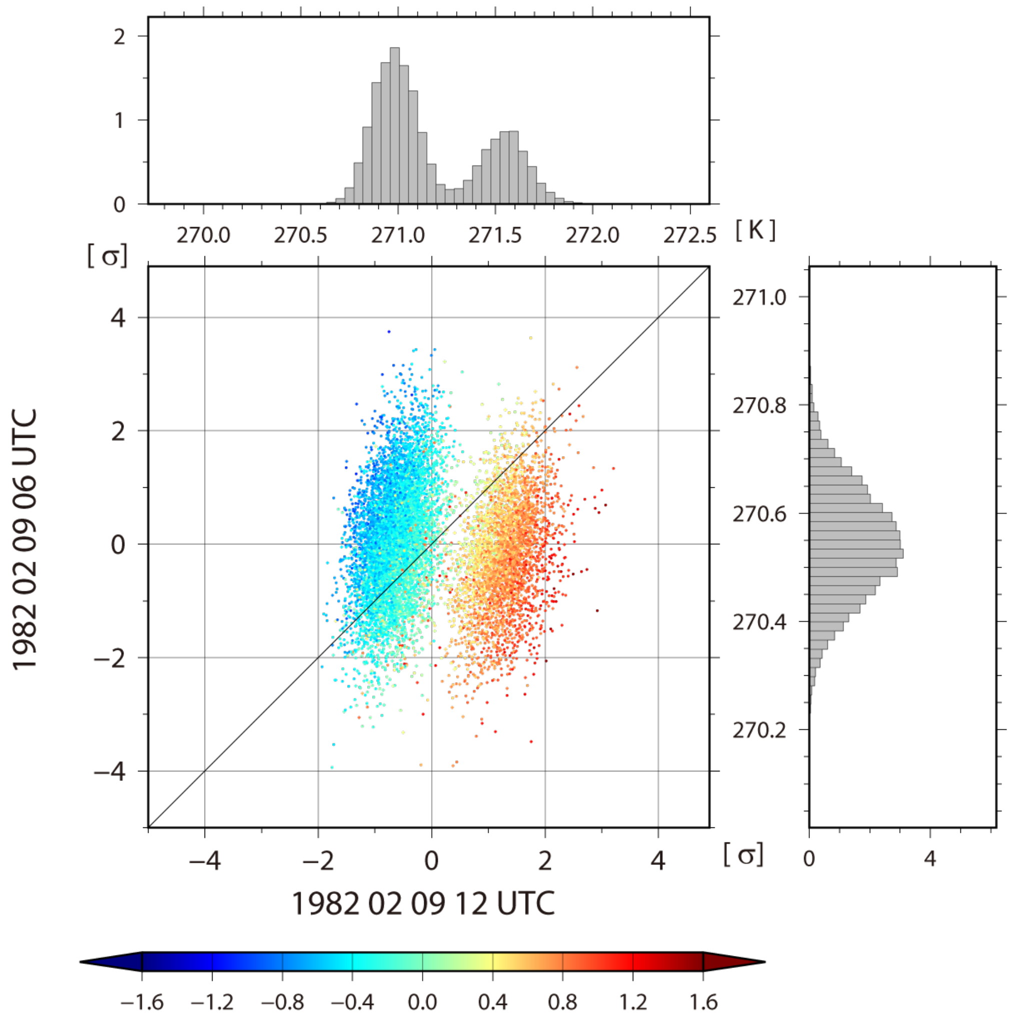




Figure 11: Scatter diagram of 0600 UTC versus 1200 UTC 9 February for the background temperature at the fourth model level at 1.856°N, 168.7°E. The colors show $d\theta_e' = (d\theta_{e\,1200\,UTC} - d\theta_{e\,0600\,UTC}) - (d\bar{\theta}_{e\,1200\,UTC} - d\bar{\theta}_{e\,0600\,UTC})$. The histograms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

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018

Nonlinear Processes in Geophysics
Discussions

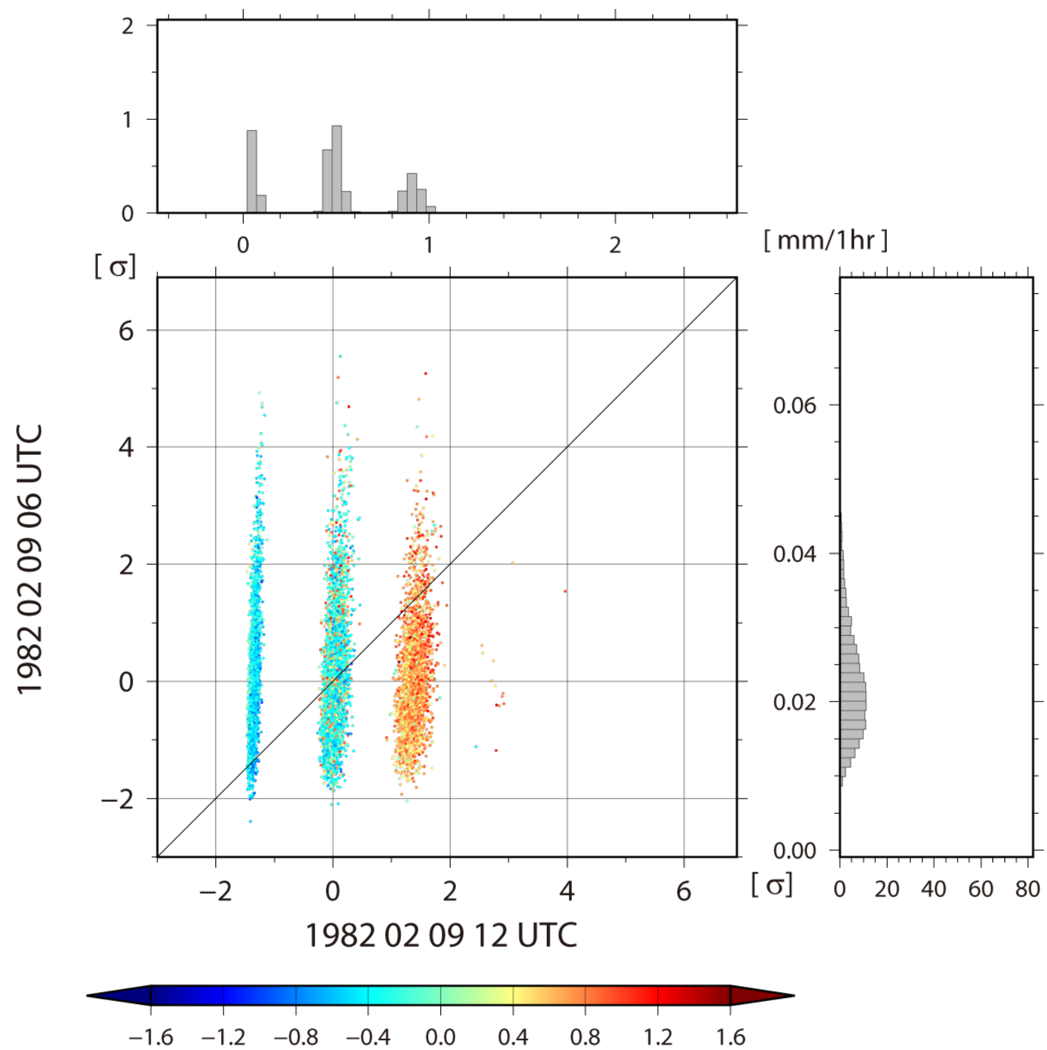
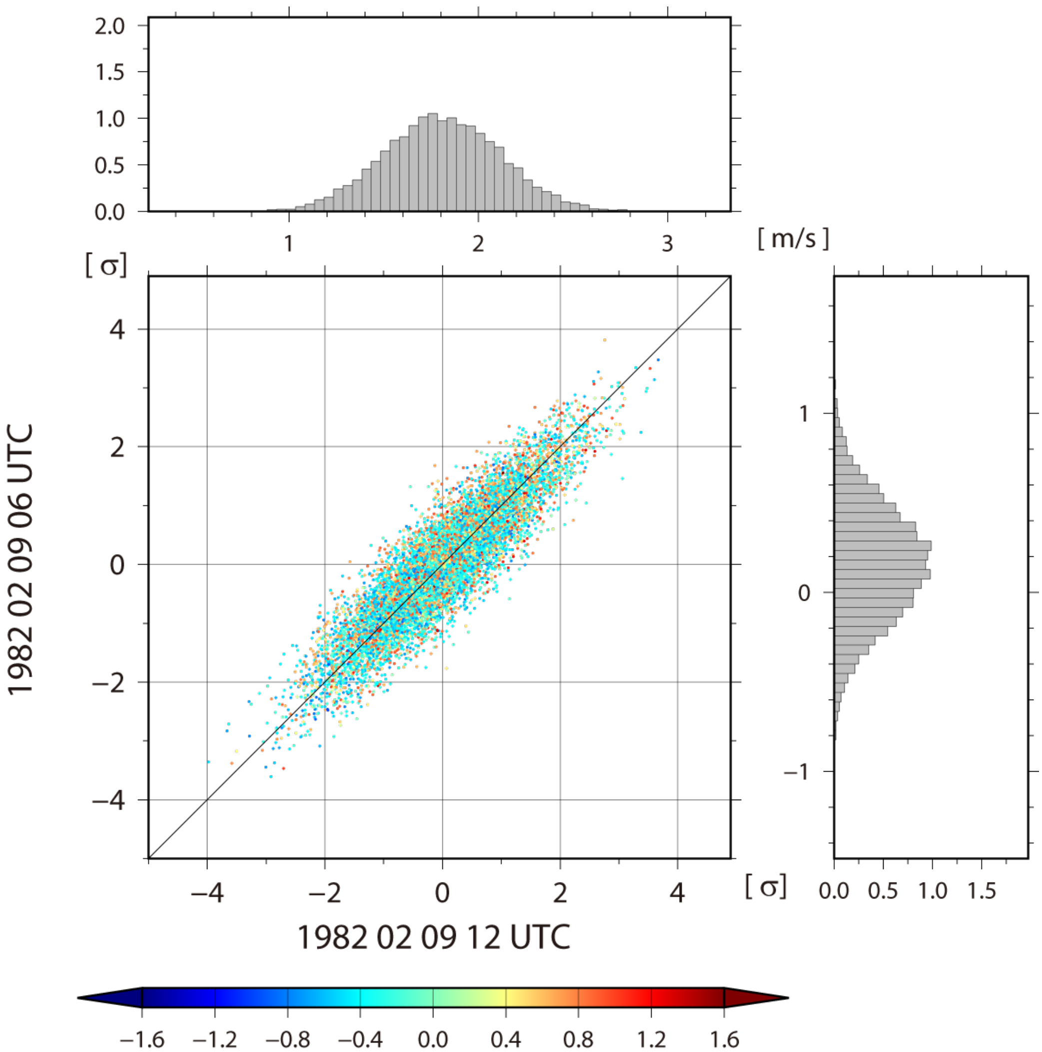



Figure 12: As in Fig. 10, but background precipitation at 0600 UTC versus background temperature at 1200 UTC 9 February.

Nonlinear Processes in Geophysics
Discussions




505

506 Figure 13: As in Fig. 10, but background zonal wind at 0600 UTC versus background temperature

507 at 1200 UTC 9 February.

508

Nonlinear Processes in Geophysics
Discussions
Open Access

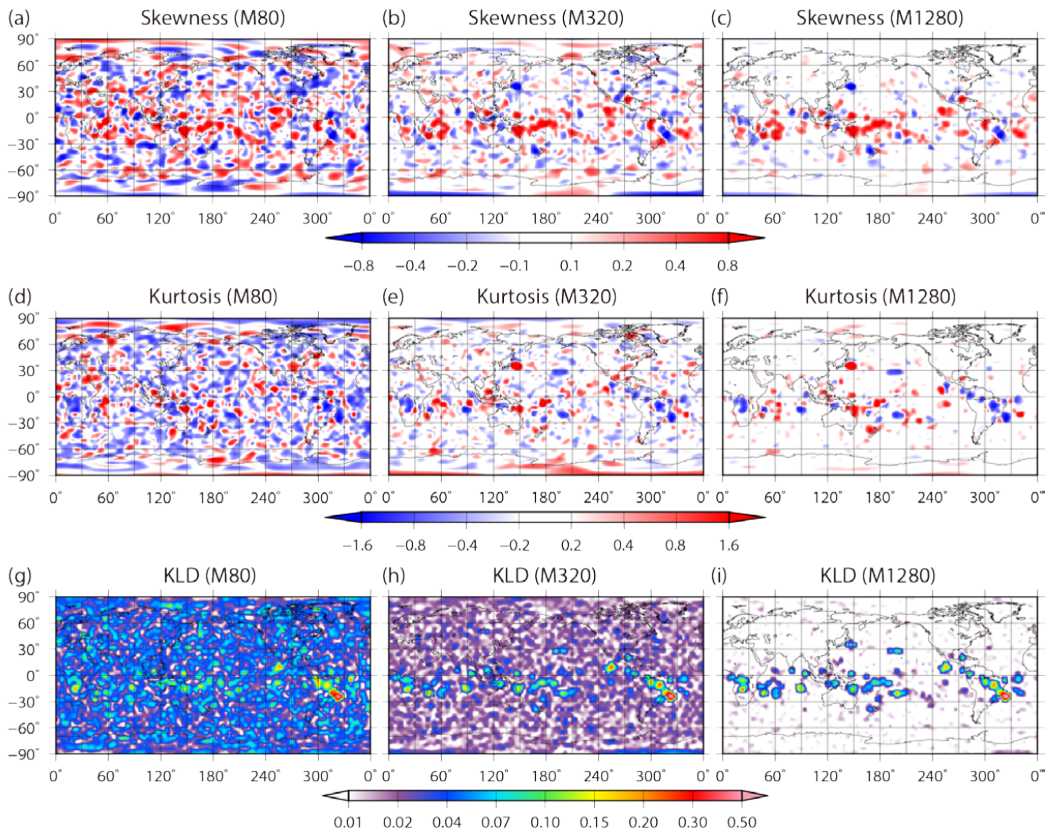



Figure 14: Spatial distributions of (a-c) skewness, (d-f) kurtosis, and (g-i) KL divergence for temperature at the fourth model level (~500 hPa) at 0600 UTC 22 February. The left, center, and right columns show 80, 320, and 1280 subsamples from 10240 members, respectively.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018


Nonlinear Processes in Geophysics
Discussions
Open Access

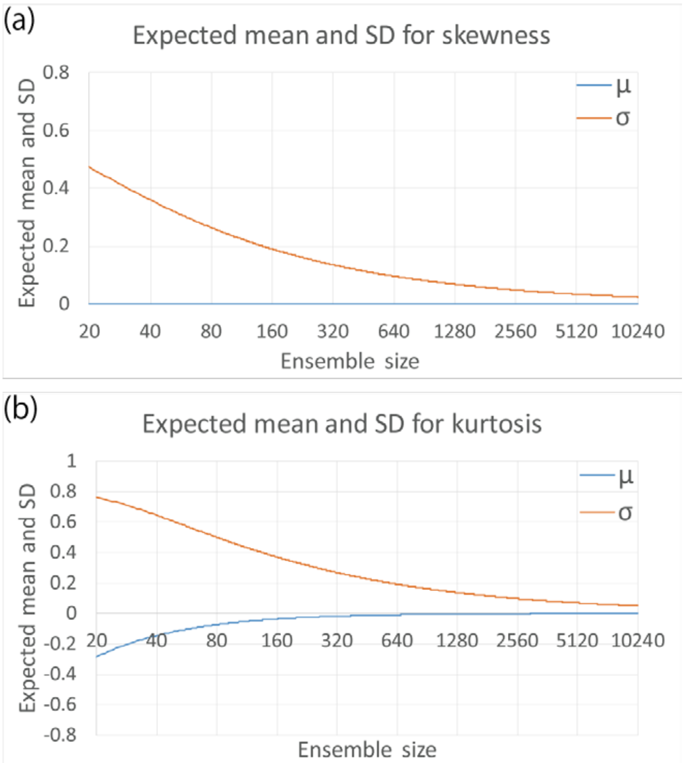



Figure 15: Expected means and standard deviations of (a) skewness $\beta_1^{1/2}$ and (b) kurtosis $\beta_2$ with increasing ensemble size up to 10240. Blue and orange curves show the mean and standard deviation, respectively.

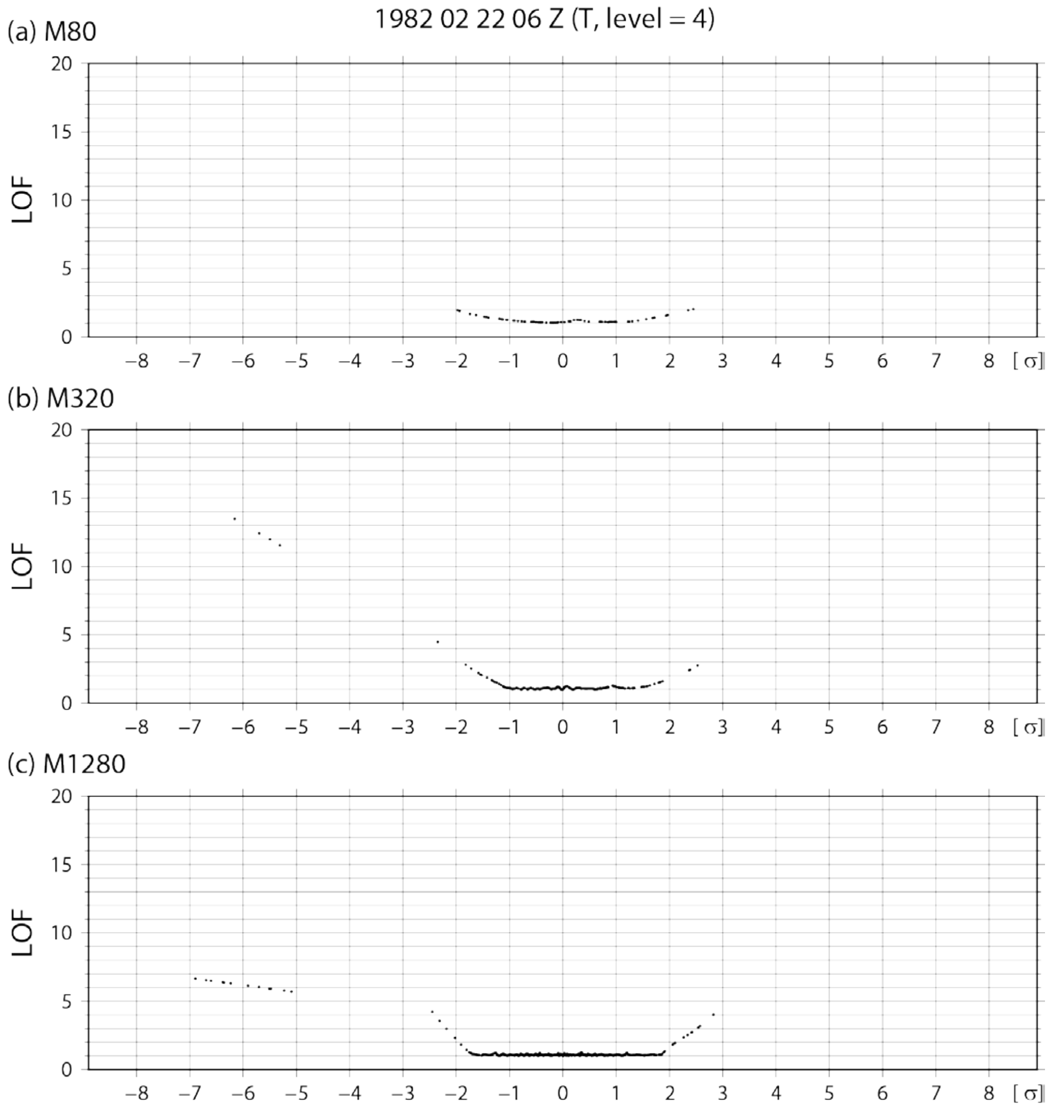



Figure 16: As in Fig. 5b, but for ensemble size (a) 80, (b) 320, and (c) 1280.

Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-52
Manuscript under review for journal Nonlin. Processes Geophys.
Discussion started: 12 November 2018

Nonlinear Processes in Geophysics
Open Access
Discussions

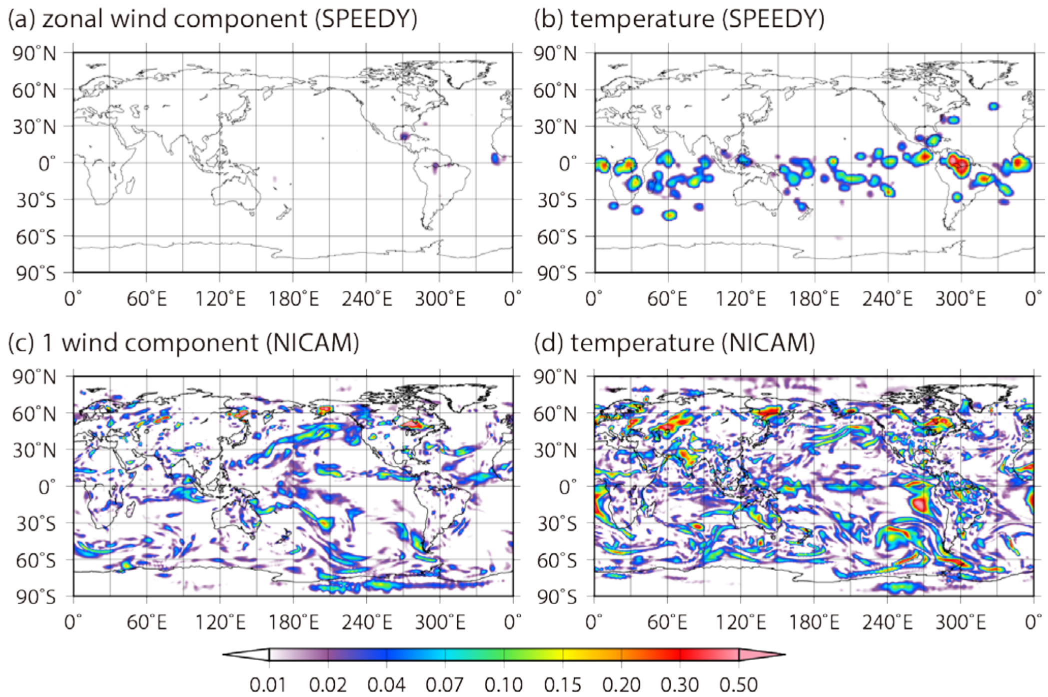



Figure 17: Spatial distributions of background KL divergence for SPEEDY model and NICAM. Upper panels show (a) zonal wind and (b) temperature at the second model level (~850 hPa) for the SPEEDY model at 0000 UTC 1 March. Bottom panels show (c) one of three horizontal wind components and (d) temperature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