Nonlinear Processes in Geophysics

Open Access

Discussions

Identifying non-normal and lognormal characteristics of temperature, mixing ratio, surface pressure, and wind for data assimilation systems

A. J. Kliewer, S. J. Fletcher, A. S. Jones, and J. M. Forsythe

Cooperative Institute for Research in the Atmosphere, Colorado State University, 1375 Campus Delivery, Fort Collins, CO 80523-1375, USA

Correspondence to: A. J. Kliewer (anton.kliewer@colostate.edu)

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Abstract

Data assimilation systems and retrieval systems that are based upon a maximum likelihood estimation, many of which are in operational use, rely on the assumption that all of the errors and variables involved follow a normal distribution. This work develops a series of statistical tests to show that mixing ratio, temperature, wind and surface pressure follow non-normal, or in fact, lognormal distributions thus impacting the design-basis of many operational data assimilation and retrieval systems. For this study one year of Global Forecast System 00:00 UTC 6 h forecast were analyzed using statistical hypothesis tests. The motivation of this work is to identify the need to resolve whether or not the assumption of normality is valid and to give guidance for where and when a data assimilation system or a retrieval system needs to adapt its cost function to the mixed normal-lognormal distribution-based Bayesian model. The statistical methods of detection are based upon Shapiro–Wilk, Jarque–Bera and a χ^2 test, and a new composite indicator using all three measures. Another method of detection fits distributions to the temporal-based histograms of temperature, mixing ratio, and wind. The conclusion of this work is that there are persistent areas, times, and vertical levels where the normal assumption is not valid, and that the lognormal distribution-based Bayesian model is observationally justified to minimize the error for these conditions. The results herein suggest that comprehensive statistical climatologies may need to be developed to capture the non-normal traits of the 6 h forecast.

1 Introduction

It has been documented several times that there are variables in the atmosphere that come from non-normal distributions (Biondini, 1976; López,1977; Mielke et al.,1977; Toth and Szentimrey, 1990; Sauvageot, 1994; Yang and Pierrehumbert, 1994; Miles et al., 2000; O'Neill et al., 2000; Harmel et al., 2002; Stephens et al., 2002; Foster and Bevis, 2003; Zhang et al., 2003; Cho et al., 2004; Sengupta et al., 2004; Foster

et al., 2006; Perron and Sura, 2013). It is shown in Fletcher (2010), that atmospheric variables may come from different probability distributions depending on the season. If this is the case then the variables' inherent distributions could also be conditioned on large-scale climatic dynamics.

If a testing procedure can be established that determines the nature of a variable, then an appropriate analysis scheme may be chosen to suit the variable. Many numerical weather prediction centers include some form of variational data assimilation, or Kalman filter, for their analyses and forecasting scheme, that are dependent on a normal distribution assumption for the error description. These centers include the Met Office (Rawlins et al., 2000), the European Centre for Medium-Range Weather Forecasts (Rabier et al., 2000), Météo-France (Fischer et al., 2005), Meteorological Service of Canada (Gauthier et al., 2007), the Naval Research Laboratory (NRL) Atmospheric Variational Data Assimilation System-Accelerated Representer (Rosmond and Xu, 2006), and the National Centers for Environmental Predication's Gridpoint Statistical Interpolation (Kleist et al., 2009). For more thorough reviews of variational data assimilation see Fletcher (2010) and Fletcher and Jones (2014).

In addition to operational data assimilation systems, the normal distribution assumption for the modeling of errors is also made for satellite retrieval systems, for example in the National Oceanic and Atmospheric Administration's Microwave Integrated Retrieval System (MiRS) (Boukabara et al., 2011), but where a logarithmic transform is used to convert a lognormally distributed variable into a normally distributed variable. This transform approach is also used in the Canadian Middle Atmosphere Model to make the state more normally-distributed (Polavarapu et al., 2005). While moment statistics have been used to analyze atmospheric variables (Perron and Sura, 2013), as of this writing the authors are unaware of any testing procedure attempting to classify the statistical framework of mixing ratio, temperature, wind and surface pressure.

There have been previous studies that have shown that variables including precipitation (Biondini, 1976; Mielke et al., 1977; Sauvageot, 1994; Cho et al., 2004), total precipitable water (Foster and Bevis, 2003; Foster et al., 2006), extreme temperatures (Toth and Szentimrey, 1990; Harmel et al., 2002), cloud and radar echo populations (López, 1997) cloud droplet size (Miles et al., 2000), liquid water path (Sengupta et al., 2004; Stephens et al., 2002), aerosol optical depth (O'Neill et al., 2000), tropical water vapor (Zhang et al., 2003), and relative humidity (Yang and Pierrehumbert, 1994) do not conform to a normal distribution to describe their behavior. A climatology of nine variables' distributional characteristics is analyzed in Perron and Sura (2013). Some of the studies considered spatial data while others used time-series data. These studies used a variety of techniques to quantify the nature of these distributions including probability density-fitting via moment calculations, χ^2 goodness-of-fit tests, moment statistics, the Shapiro–Wilk tests, and simply plotting histograms with prominent non-normal distribution features.

Across a variety of disciplines it is often convenient, and somewhat innocuous, to treat measured variables as normally distributed in nature. This can misrepresent the inherent summary statistics due to a loss of information (e.g., lack of higher statistical moment information), and can be harmful within certain applications of the data. If a model, or algorithm, incorrectly assumes that a random variable is normally distributed then the properties of this distribution may skew its output.

A variable's probability distribution dictates the probabilistic solution that is found when using data assimilation techniques. In 3-D variational assimilation the cost function is given by

$$J(x) = \frac{1}{2}(x - x_{\mathrm{b}})^{\mathsf{T}}\mathbf{B}^{-1}(x - x_{\mathrm{b}}) + \frac{1}{2}(y - h(x))^{\mathsf{T}}\mathbf{R}^{-1}(y - h(x)) \tag{1}$$

where $\mathbf{B}$ is the background error covariance matrix, $\mathbf{R}$ is the observational error covariance matrix, y is the vector of observations, and $h(x)$ is the non-linear observation operator. The background component of Eq. (1) includes the difference between the minimizing solution x and the background distribution x_{b}. If both x

Non-normal atmospheric variables

A. J. Kliewer et al.

and x_b are assumed to be indepedent normally-distributed, then the difference of these variables is also a normally-distributed random variable. If these variables are not normally-distributed, or moreover lognormally-distributed, then the difference is not a normally-distributed random variable (Fletcher, 2010), however the ratio of the variables is a lognormally-distributed random variable (Casella and Berger, 2002). It is an implicit model assumption that the solution x and the background x_b come from the same probability distribution. It should also be noted that if the "errors" are assumed to be unbiased then $\mu_{true} = \mu_b$, which means that the expected value of the minimizing solution x and background x_b are centered at the same point but have different spread (variance) and therefore skewness. It has been previously observed that that mixing ratio errors are not normally distributed (Dee and da Silva, 2003) and in fact are lognormally distributed (Daley and Barker, 2001).

Biases could also be introduced in data assimilation and retrieval systems that assume variables, and hence their errors, are normally distributed when they actually follow a non-normal distribution in nature. A clear example of where this can be problematic is if a computed value is physically impossible, such as relative humidity taking a negative value. This dubious value may be incorrectly incorporated into the analyses, or reset to a lower bound near zero. In either case this is certainly less desirable than solving for the correct value using an appropriate scheme that incorporates its correct underlying probability distribution. Recently, mixed normal-lognormal variational data assimilation methods have been developed in 3-D (Fletchera and Zupanski, 2006a, b, 2007), and 4-D in Fletcher (2010). These initial full field formulations were not consistent with the current operational incremental configurations. However, a derivation and testing of a mixed multiplicative and additive incremental 3-D- and 4-D-VAR for a control vector that contains both normal and lognormally distributed variables is presented in Fletcher and Jones (2014).

Evidence of how an assimilation scheme improves based on the distribution of the observational errors is shown in Fletcher and Jones (2014). Using the Lorenz '63 chaotic model the authors show that a lognormal-based cost function performs better

than the current normal formulation given lognormal errors. Those conclusions result from testing observations of varying accuracy, sparseness in time, and over different window lengths.

Given that there is now a mathematical framework for assimilating mixed normal-lognormally distributed variables/errors, techniques are needed that can inform the user of a mixed system when to switch between a full normal distribution-based version or a mixed normal-lognormal-based version to optimize the performance of the system and to make it consistent with the "current" observed probabilistic behavior.

Therefore, the motivation of this work is to design a set of tests that can be performed offline between cycles or windows such that the configuration for the approximation for the background error covariance matrix, cost function, Jacobian and approximations to the Hessian, if used, can be ready for the next minimization step.

Given the motivation to detect a non-normal, specifically a lognormal signal, we use 1° resolution data from the National Oceanic and Atmospheric Administration (NOAA) Global Forecast System (GFS) 00:00 UTC 6 h forecast between 1 January 2005 and 31 December 2005 defined on a 181×360 grid. The forecasts, which are the GFS outputs, at each grid point form the time series. The data is analyzed over the entire year as well as on a "seasonal" basis by considering 3 months at a time (January–March, April–June, July–September, and October–December). The sample sizes are consistent with the suggested size from Croarkin and Tobias (1999) of $2n^{\frac{2}{5}}$, where n is the total number of observations available. The chosen variables include mixing ratio, temperature, surface pressure and wind at levels 100, 200, 300, 500, 700, 850, and 1000 hPa.

While there are transformation techniques employed by operational centers for moisture (Bocquet et al., 2010), the Navy Operational Global Atmospheric Prediction System (NOGAPS) previously used the logarithm of specific humidity (Eckermann et al., 2004), which is equivalent to mixing ratio (Dee and da Silva, 2003) analyzed in this study. In Fletcher and Zupanski (2007) it is shown that a logarithmic transformation

finds the median in multivariate lognormal space, which is positively biased relative to the mode, or the most likely state.

In this work we propose using easily calculable statistics and hypothesis testing to show that these variables described above show strong evidence of a non-normal nature, or more specifically, a lognormal behavior. The hypothesis tests considered in this paper include the Jarque–Bera, Shapiro–Wilk, and χ^2 goodness-of-fit. In addition, a composite hypothesis test is proposed that includes all of the decisions made by the aforementioned tests. Such tests are a requirement of advanced methods (Fletcher and Zupanski, 2007; Fletcher, 2010; Song et al., 2012; Fletcher and Jones, 2014) that are able to use multiple probability models.

The format of the remainder of this paper proceeds as follows: Sect. 2 describes the formulation of the hypothesis tests as well as the test statistics. In Sect. 3 results of these tests are presented. In Sect. 4 conclusions and a discussion of the results of Sect. 3 are presented.

2   Statistical methods

In this section the statistical methods that are used to detect a non-normal distribution signal are presented along with tests to see if the distribution is a lognormal distribution. The random sample $x_1, \ldots, x_n \in X$ of independent and identically distributed (iid) observations is taken from the GFS data for each of the hypothesis tests that all rely on a significance level of $\alpha = 0.01$. This value of α indicates a 99 % confidence-level in the results of the testing procedures.

The samples' autocorrelation has been checked in order to verify the iid assumption for the hypothesis tests. While there is some autorcorrelation in the samples, we attempt to minimize its effect by choosing such a small α. Histograms of the data are also presented in order to verify the validity of the results of the hypothesis tests as well. The iid assumption on any data set found in nature is difficult to assert and it is

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

also noted that many methods, including the National Meteorological Center (NMC) method (Parrish and Derber, 1992), make no correction for autocorrelation.

2.1 Hypotheses

For the Shapiro–Wilk and the Jarque–Bera tests (Hain, 2010) the following hypotheses are defined, with $-\infty < \mu < \infty$ and $0 < \sigma < \infty$,

$$H_0 : p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty \tag{2}$$

$$H_a : p_X(x) \neq \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

In all subsequent presentations of results a returned value of 0 in a hypothesis test indicates that the null hypothesis cannot be rejected at the α-level. A value of 1 indicates that the null hypothesis is rejected in favor of the alternative hypothesis. The subtlety of this framework cannot be overstated in that while the data may in fact originate from a non-normal distribution there may not be enough evidence in the data to support the claim that it is not normally distributed and therefore the result of the hypothesis test will be 0. It is assumed that the null hypothesis is true prior to the test, thus putting the burden of proof on the alternative hypothesis, with the choice of $\alpha = 0.01$ indicating that the testing procedures are very conservative. While one of the aims of this paper is to investigate the possibility of mixing ratio, temperature, wind and surface pressure following a lognormal distribution, this conclusion is not possible with the previous hypotheses. Therefore a χ^2 goodness-of-fit test has the following

hypotheses, with $-\infty < \mu < \infty$ and $0 < \sigma < \infty$,

$$H_0 : p_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \qquad x > 0 \qquad (3)$$

$$H_a : p_X(x) \neq \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \qquad x > 0.$$

In an attempt to combine both sets of hypotheses a new "composite test" is defined. In this test if both the Shapiro–Wilk and the Jarque–Bera tests reject H_0 in favor of H_a, and the χ^2 test fails to reject H_0, then a value of 1 is returned, otherwise the result is 0. This is meant to be a very strict test of the data not coming from a normal distribution but in fact that the data is from a lognormal distribution.

As opposed to reporting the skewness and kurtosis of a particular time-series as in Perron and Sura (2013), this information is used to make a decision about the distribution. While the structure of a hypothesis test includes a preconception about the data, multiple tests are combined simultaneously to test both directions of the normality assumption. This design ensures that the data truly is lognormally distributed without a false positive. The authors are not aware of this technique having been previously applied.

2.2 Shapiro–Wilk

Let $x_{(1)}, \ldots, x_{(n)}$ be the order statistics of the random variable X and $\overline{x}$ the sample mean, where the order statistic of rank k is the kth smallest value in X, denoted by $x_{(k)}$. Define the vector $\boldsymbol{m} = (m_1, \ldots, m_n)^\mathsf{T}$, where $m_1, \ldots, m_n$ are the associated expected values of $x_{(i)}$, and let $\mathbf{V}$ be the covariance matrix of the order statistics. The expected value and covariance matrix for a random variable X with probability density function $f(x)$ are

given by

$$E(\mathbf{X}) = \int\limits_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x \tag{4}$$

$$V = E\left[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^{\mathsf{T}}\right]. \tag{5}$$

Then the Shapiro–Wilk (SW) test statistic is given by

$$\mathrm{SW} = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}, \tag{6}$$

where

$$(a_1, \ldots, a_n) = \frac{\boldsymbol{m}^{\mathsf{T}}\mathbf{V}^{-1}}{\left(\boldsymbol{m}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{V}^{-1}\boldsymbol{m}\right)^{\frac{1}{2}}}. \tag{7}$$

A thorough mathematical explanation of this statistic is presented in Hain (2010). Razali and Wah (2011) has found that the Shapiro–Wilk test outperforms in power the Kolmogorov–Smirnov, Lilliefors, and Anderson–Darling tests for both symmetric and non-symmetric distributions based on sample size. The power of a test is the probability of not committing a Type II error, which occurs when H_0 is false but is incorrectly not rejected (Casella and Berger, 2002).

2.3 Jarque–Bera

Clear differences between the normal and lognormal distributions include skewness and kurtosis. Skewness essentially determines the asymmetry of a distribution. This statistic can be positive, negative or zero and is the third moment of a random variables' probability distribution. Kurtosis, the fourth moment, measures how peaked

the distribution is. Descriptions of these statistics can be found in Casella and Berger (2002). The Jarque–Bera test combines these statistics to determine their goodness-of-fit to a normal distribution. If the distribution is normal, then asymptotically the Jarque–Bera (JB) test statistic has a χ^2 distribution with two degrees of freedom and is given by

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right),\qquad(8)$$

with the third and fourth moments given by

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^{\frac{3}{2}}},\qquad(9)$$

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^{2}}.\qquad(10)$$

2.4 Chi-squared

With the null hypothesis of the chi-squared test being that the data come from a lognormal distribution, the test statistic compares expected, E_i, vs. observed, O_i, observations in k bins of data. The expected frequency, for each bin, is given by

$$E_i = n\left(F\left(Y_u\right) - F\left(Y_l\right)\right),\qquad(11)$$

where F is the cumulative distribution function for the lognormal distribution and Y_u and Y_l are the upper and lower limits for class i and n is the sample size. The statistic, which is compared against the χ^2 distribution, is given by

$$\chi^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}.\qquad(12)$$

Much more can be said about these hypothesis tests but that is outside of the scope of this paper. Those details are left out in lieu of the application results as applied to the GFS data.

2.5 Distribution fitting

The normal and lognormal probability density functions are fitted to the data using the maximum likelihood technique. For an independent and identically distributed sample $X_1, \ldots, X_n$ with probability density $f(\mathbf{x}|\theta_1, \ldots, \theta_k)$, the likelihood function is defined by

$$L\left(\theta|x\right) = L\left(\theta_1, \ldots, \theta_k|x_1, \ldots, x_k\right) = \prod_{i=1}^{n} f\left(x_i|\theta_1, \ldots, \theta_k\right). \tag{13}$$

For each sample point x the likelihood function is maximized as a function of θ. A thorough explanation of this procedure can be found in Casella and Berger (2002).

3 Results

The results of the time-series hypothesis tests for mixing ratio and temperature resulted in numerous figures and data plots displaying the non-normal and lognormal nature of the GFS data. An overview of these results is presented along with a more detailed analysis of specific points of interest. Instead of presenting the results of the Shapiro–Wilk, Jarque–Bera, and Chi-squared tests only the results of the Composite Test are shown which incorporate all of the results simultaneously.

For each point of the GFS data, an forecast from each day between 1 January 2005 through 31 December 2005 makes up the random variable X for one year. This data is also broken down into four "seasons," i.e. 1 January 2005 through 31 March 2005 (denoted as JFM in all figures), 1 April 2005 through 30 June 2005 (denoted AMJ), 1 July 2005 through 30 September 2005 (denoted JAS), and 1 October 2005 through 31 December 2005 (denoted OND).

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

3.1 Mixing ratio

A tabulated view of all of the tests results can be seen in Fig. 1. Frequencies depict how often the Shapiro–Wilk and Jarque–Bera tests reject H_0, the Chi-squared failed to reject H_0, and when these results coincided for each point of the GFS. Therefore the Composite test cannot have a larger value than any one of the individual tests. For example, the entire year of forecasts of mixing ratio at 300 hPa has almost 99 % of points coming from a non-normal distribution as concluded by the Shapiro–Wilk and Jarque–Bera tests, and almost 29 % of points cannot be determined to not come from lognormal distribution as per the Chi-squared test. Therefore the composite test concludes that almost 29 % are lognormally distributed. This chart demonstrates that there is significant occurance of the non-normal distribution behavior, but not necessarily lognormal behavior as determined by the Chi-squared test. Choice of $\alpha = 0.01$ dictates these results and can be adjusted depending on the user's desired level of confidence.

Figure 2 shows the results of the composite test at 300 hPa. In this and all subsequent figures red areas indicate a positive result of the composite test, i.e. the Shapiro–Wilk and Jarque–Bera rejected the hypothesis that the data come from a normal distribution and the chi-squared test failed to reject the hypothesis the data come from a lognormal distribution. Blue areas in these figures indicate that at least one of these conclusions is not met for the hypothesis tests. With the composite test it is easy to see when all of the tests agree that the data comes from a lognormal (red) as opposed to a non-lognormal distribution (blue). It is interesting to note how the data changes over the course of the year as well as when the data is taken as a whole for 2005. Since the areas in red are not randomly scattered, coherent physical processes must be at work to sustain the statistical properties of the mixing ratio. Similarly, Fig. 3 shows the results of the composite test at 500 hPa for each time domain. Note that the first two time domains of 2005 have the largest coverage of lognormally-distributed data.

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

|◄ | ►|

◄ | ►

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

To see what a sample of the data actually looks like consider Fig. 4. This data is at 300 hPa located in the North Atlantic off the Canadian coast. For 2005 as a whole as well as each season the composite test returned a positive result for this location. With the fitted probability distributions it is clear that the lognormal distribution is a better fit for the mode of the data and also captures its skewness. Conversely, the fitted normal distribution misses the mode, attempts to smooth out the data, and includes substantial probabilites for values below zero which is physically impossible for mixing ratio. Cold dry air extrusions into this region could very well be driving this statistical behavior.

Another location of interest which experiences significant continental air masses (Trewartha and Horn, 1971) is in central North America where tornadoes frequently develop. Figure 5 shows the data and probability fits at 300 hPa. Once again this is an instance where 2005 and each season passes the composite test. The lognormal distribution tightly fits data again whereas a characteristic of the normal distribution is that it is centered around the mean of the data, which is not necessarily the location parameter of choice for a skewed distribution. As a result, a symmetric curve is placed at the mean which misses the major characteristics of the data.

Figure 6 shows the data and distribution fits for a point in the tropical cyclone formation region in the North Atlantic at 500 hPa. In this instance, the composite test passes for each season but not for the year even though the histogram resembles a lognormal distribution. The reason for this speaks to the conservative nature of the tests. With $\alpha = 0.01$ and the sample size of $n = 363$, there must be overwhelming evidence for all of the tests, and therefore the composite test, to conclude the data's true statistical signal. Similar to Fig. 4, the lognormal fit clearly captures the nature of the data better than the normal distribution.

For a location near Japan at 850 hPa the composite test correctly concludes that the data does not follow a lognormal distribution as shown in Fig. 7. Here the data is either somewhat symmetric or is bi-modal. For the January–February–March months the normal fit is somewhat better than the lognormal. However, for the entire year the

normal fit misses both modes entirely and gives maximum probability to less observed values.

Closer inspection of many more vertical levels and locations could be shown but are omitted due to limitations of space.

3.2 Temperature

Similar to Fig. 1 for mixing ratio, statistical test results are presented for temperature in Fig. 10. It is clear that the composite test concludes that the non-normal and lognormal signals are seen to be much less pronounced for temperature than for mixing ratio. However there are still numerous occurances as determined by the strict hypothesis tests. Inspection of the composite test results for 500 and 700 hPa can be seen in Figs. 8 and 9. In these images it is clear that the lower tropics are more likely to have lognormally-distributed temperature data.

By looking at the results of the Shapiro–Wilk and Jarque–Bera tests, there are occurances where the temperature data is seen to come from a non-normal distribution. There are 77 points out of 65 160 where the data for all of 2005 and each season is not normally distributed, i.e. the null hypothesis is rejected for these tests on all time domains. All but one of these points are in the Southern Hemispere, with a majority of points falling between 500 and 1000 mb. in the Southern Indian Ocean. Results shown in Figs. 11 and 12 contain examples of this occurring in the Indian Ocean and near Japan respectively. Note how the data is either bi-modal, positively- or negatively-skewed, or even resembles a uniform distribution. In addition there are numerous points, across all pressure levels, where the data for one or more "seasons" is not normaly-distributed.

3.3 Surface pressure

While surface pressure is a positive definite random variable, the chi squared test indicated no instances of lognormal behavior. This is a result of the data typically being right-skewed if the normal assumption is rejected.

While non-normal behavior is not as prevalent in surface pressure as in mixing ratio, the frequency can be seen in Fig. 13. Here the composite test indicates the frequency that the Jarque–Bera and Shapiro–Wilk reject the null hypothesis, omitting the Chi squared test. Spatial coverage of the composite test is shown in Fig. 14.

An interesting presentation of the number of seasons where the normality assumption is rejected by the composite test is shown in Fig. 15. Here, areas over the ocean are seen more often to have non-normally distributed surface pressure than over land.

3.4 Wind

Since the GFS wind data is not a positive definite random variable, the lognormal distribution is not a viable candidate to capture its shape or spread. Therefore, for wind, the composite test now reports when both the Shapiro–Wilk and Jarque–Bera tests simultaneously reject the null hypothesis that the data comes from a normal distribution. Since a much more thorough review of the probability distributions of wind has been conducted by Carta et al. (2009), a brief inclusion of the results is presented here, which corroborate the non-normal behavior of wind that has been previously observed.

Figures 16 and 17 show the frequency that each test rejected normality as well as where they overlap in the composite test for the u and v wind components. It is clear that for almost every time domain, the vertical level with the least percentage of non-normal behavior ("most normally-distributed") is at 500 hPa. Also of interest are the differences in the wind analyses at 1000 hPa, which clearly show that the u component is more likely to be non-normal. This can be seen spatially in Fig. 18.

Closer inspection of the nature of the skewed and bi-modal behavior of u can be seen in Fig. 19. For each time domain at 850 hPa, the normal assumption is rejected by the composite test. The normal distribution misses the mode of the 0 h forecast and the presence of values less than zero prevent the fit of a lognormal distribution.

Given these results for mixing ratio, temperature, surface pressure, and wind, a real-time detection method may include a moving-average that includes the last t number of forecasts, where this value t could be user-defined to specify a certain power for the hypothesis tests. Another method may involve including data available in the current season in order to make a determination of whether to assume the variable is normally- or lognormally-distributed. As demonstrated in this paper the hypothesis tests are robust for a t smaller than one year.

In this section different variables, vertical levels, time domains, and locations have been presented demonstrating non-normal or logormal (or neither) behavior. Given the prevalence of non-normally distributed random variables the necessity of checking what the data looks like has been demonstrated.

4 Conclusions and discussion

Since mixing ratio and temperature have been shown to be non-normally distributed and in many cases appear to be lognormally distributed, 3-D- and 4-D-VAR data assimilation schemes that include lognormal cost functions for both the observations and the apriori background may be required for more accurate results. This would have implication on the forecast skill of a DA system, or for a retrieval system, as the analysis state from the minimization of the mixed distribution cost function should be consistent with the probabilistic behavior of the true state. The normal assumption, while convenient and easily adaptable, may need to be more carefully considered in light of these results.

While it is true that a lognormal distribution with a small variance looks very similar to a normal distribution, the detection methods used in this paper attempt to operationally

handle large amounts of data similar to the resolution of an inner loop in incremental data assimilation schemes. It is in this end that these statistical procedures have been demonstrated in order to understand the true nature of atmospheric variables.

The time-series data clearly indicates data for mixing ratio and temperature will follow a lognormal distribution in certain areas. These results give light to the fact that the normal distribution assumption is not a valid assumption for the basis of the data assimilation and variational based retrieval systems and suggests that more research is needed to study the impact of assuming a normal distribution fit on forecast skills, variational observational quality control as well as the gross error check (Lorenc and Hammon, 1988).

Therefore this work suggests that statistical climatology tests need to be developed on a seasonal, or possibly a monthly basis, as the distributions that are found for specific variables indicate which distribution's cost function should be used in the assimilation schemes as a function of space and time. Ideally a real-time decision of how the data is statistically structured would be determined, ensuring that the correct scheme is chosen. In either case, it is the goal that an objective decision methodology be available for an appropriate scheme based on the nature of the data. The choice under what observational conditions to apply alternative Baysian models is now made as an objective decision through the procedure used and demonstrated in this work.

Future work can consider longer time-series, more vertical levels, other atmospheric variables such as column water vapor when a boundary layer cloud is present as seen in Fletcher (2010), and other statistical methods including the Akaike information criterion (Akaike, 1974). The possible future benefit of the Akaike information criterion (AIC) is that it detects the best distribution for a random variable based on information theory which could then give guidance for what other distributions need to be included in the variational cost function. AIC balances the goodness-of-fit of a distribution while minimizing the number of model parameters.

It has been shown in Fletcher and Jones (2014) that there is a negative impact on the performance of a normal distribution only incremental 4-D-VAR when lognormal

Non-normal atmospheric variables

A. J. Kliewer et al.

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

◀ | ▶|

◀ | ▶

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

forecasts are assimilated. However, when the same observations were assimilated in a lognormal-based incremental 4-D-VAR, then there was no negative impact on the analysis error. Therefore, the need to determine which distribution the observations and their errors come from is important to minimize the impact of these errors on the analysis of a DA system and the subsequent forecast. In this paper methodologies have been developed and tested with the 2005 GFS 00:00 UTC 6 h forecast and it has been shown that there are lognormal signals in the forecasts. This therefore suggests a need for statistical climatologies to be developed and for these climatologies to also be linked in near real-time with the data assimilation and retrieval systems.

Appendix: Distribution of errors

Let η be the background error component of the 3-D cost function Eq. (1) given in Sect. 1, i.e. let

$$\eta = (x - x_b)^\top .\qquad\qquad(A1)$$

Without loss of generality consider the univariate case. For a random variable X with a cumulative density function F_X, the moment generating function is defined by $M_X(t) = E[e^{tX}]$. The moment generating function for a normal random variable with mean μ and variance σ^2 is given by

$$M(t) = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} .\qquad\qquad(A2)$$

Since this equation is an exponential, the sum Z of two independently distributed normal random variables is also a normal random variable. That is, if $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$, then $X + Y = Z \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$. The uniqueness theorem states that if two random variables have the same momement generating function, then they

have the same probability distributions. Clearly,

$$M_Z(t) = E[e^{tZ}] = E[e^{t(X+Y)}] = E[e^{tX}]E[e^{tY}] \tag{A3}$$
$$= M_X(t)M_Y(t)$$

$$= \exp\left\{\mu_x t + \frac{\sigma_x^2 t^2}{2}\right\} \exp\left\{\mu_y t + \frac{\sigma_y^2 t^2}{2}\right\}$$

$$= \exp\left\{(\mu_x + \mu_y)t + \frac{\left(\sigma_x^2 + \sigma_y^2\right)t^2}{2}\right\}$$

$$\sim N\left(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2\right).$$

Equation (A1) can be written as

$$\eta + x_b = x. \tag{A4}$$

If it assumed that η and x_b are normally distributed random variables then it has been shown that the left hand side of Eq. (A4) is normally distributed.

Section 3 contains results that atmospheric random variables can have a non-normal, or in particular, a lognormal distribution. This would imply that the right hand side of Eq. (A4) would be the sum of a normal and a lognormal distribution. An assumption such as this for the sought after state x would be highly suspect. It is with this mathematical formulation in mind that yielded the research into mixed normal-lognormal variational data assimilation methods as well as the distributions of the assimilated variables.

References

Akaike, H: A new look at the statistical model identification, IEEE T. Automat. Contr., 19, 716–723, doi:10.1109/tac.1974.1100705, 1974.

Biondini, R.: Cloud motion and rainfall statistics, J. Appl. Meteorol., 15, 205–224, doi:10.1175/1520-0450(1976)015<0205:CMARS>2.0.CO;2, 1976.

Bocquet, M., Pires, C., and Wu, L.: Beyond Gaussian statistical modeling in geophysical data assimilation, Mon. Weather Rev., 138, 2997–3023, doi:10.1175/2010MWR3164.1, 2010.

Boukabara, S. A., Garrett, K., Chen, W., Flavio, I. S., Grassotti, C., Kongoli, C., Chen, R., Liu, Q., Yan, B., Weng, F., Ferraro, R., Kleespies, T., and Meng, H.: MiRS: An all-weather 1-DVAR satellite data assimilation and retrieval system, IEEE T. Geosci. Remote, 49, 3249–3272, doi:10.1109/tgrs.2011.2158438, 2011.

Carta, J. A., Ramírez, P., and Velázquez, S.: A review of wind speed probability distributions used in wind energy analysis: case studies in the Canary Islands, Renew. Sust. Energ. Rev., 13, 933–955, doi:10.1016/j.rser.2008.05.005, 2009.

Casella, G., and Berger, R.: Statistical Inference, Duxbury Press, Pacific Grove, CA, 2002.

Cho, H. K., Bowman, K. P., and North, G. R.: A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the Tropical Rainfall Measuring Mission, J. Appl. Meteorol., 43, 1586–1597, doi:10.1175/JAM2165.1, 2004.

Croarkin, M. and Tobias, P.: NIST/SEMATECH engineering statistics Internet handbook, available at: http://www.nist.gov/stat.handbook, last access: 26 August 2015, 1999.

Daley, R. and Barker, E.: NAVDAS: Formulation and diagnostics, Mon. Weather Rev., 129, 869–883, doi:10.1175/1520-0493(2001)129<0869:ANFAD>2.0.CO;2, 2001.

Dee, D. and da Silva, A.: The choice of variable for atmospheric moisture analysis, Mon. Weather Rev., 131, 155–171, doi:10.1175/1520-0493(2003)131<0155:TCOVFA>2.0.CO;2, 2003.

Eckermann, S. D., McCormack, J. P., Coy, L., Allen, D., Hogan, T., and Kim, Y. J.: NOGAPS-Alpha: A prototype high-altitude global NWP model, Preprint Volume, P2.6, Symposium on the 50th Anniversary of Operational Numerical Weather Prediction, American Meteorological Society, University of Maryland, College Park, MD, 14–17 June, 2004.

Fischer, C., Montmerle, T., Berre, L., Auger, L., and Ştefănescu, S. E.: An overview of the variational assimilation in the ALADIN/France numerical weather-prediction system, Q. J. Roy. Meteor. Soc., 131, 3477–3492, doi:10.1256/qj.05.115, 2005.

Non-normal atmospheric variables

A. J. Kliewer et al.

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

◄ | ►|

◄ | ►

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Fletcher, S. J.: Mixed lognormal-Gaussian four-dimensional data assimilation, Tellus A, 62, 266–187, doi:10.1111/j.1600-0870.2010.00439.x, 2010.

Fletcher, S. J. and Jones, A. S.: Multiplicative and additive incremental variational data assimilation for mixed lognormal and Gaussian errors, Mon. Weather Rev., 142, 2521–2544, doi:10.1175/MWR-D-13-00136.1, 2014.

Fletcher, S. J. and Zupanski, M.: A data assimilation method for log-normally distributed observational errors, Q. J. Roy. Meteor. Soc., 132, 2505–2519, doi:10.1256/qj.05.222, 2006a.

Fletcher, S. J. and Zupanski, M.: A hybrid normal and lognormal distribution for data assimilation, Atmos. Sci. Lett., 7, 43–46, doi:10.1002/asl.128, 2006b.

Fletcher, S. J. and Zupanski, M.: Implications and impacts of transforming lognormal variables into normal variables in VAR, Meteorol. Z., 16, 755–765, doi:10.1127/0941-2948/2007/0243, 2007.

Foster, J. and Bevis, M: Lognormal distribution of precipitable water in Hawaii, Geochem. Geophy. Geosy., 4, 1–8, doi:10.1029/2002gc000478, 2003.

Foster, J., Bevis, M., and Raymond, W.: Precipitable water and the lognormal distribution, J. Geophys. Res., 111, D15102, doi:10.1029/2005JD006731, 2006.

Gauthier, P., Tanguay, M., Laroche, S., Pellering, S., and Morneau, J.: Extension of a 3-D-Var to 4-D-Var: implementation of 4-D-Var at the Meteorological Service of Canada, Mon. Weather Rev., 135, 2339–2354, doi:10.1175/MWR3394.1, 2007.

Hain, J.: Comparison of common tests for normality, PhD thesis, Institut für Mathematik und Informatik, Julius-Maximilians-University at Würzburg, Germany, 102 pp., 2010.

Harmel, R. D., Richardson, C. W., Hanson, C. L., and Johnson, G. L.: Evaluating the adequacy of simulating maximum and minimum daily air temperature with the normal distribution, J. Appl. Meteorol., 41, 744–753, doi:10.1175/1520-0450(2002)041<0744:ETAOSM>2.0.CO;2, 2002.

Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Wu, W. S., and Lord, S.: Introduction of the GSI into the NCEP Global Data Assimilation System, Weather Forecast., 24, 1691–1705, doi:10.1175/2009WAF2222201.1, 2009.

López, R. E.: The lognormal distribution and cumulus cloud populations, Mon. Weather Rev., 135, 865–872, doi:10.1175/1520-0493(1977)105<0865:TLDACC>2.0.CO;2, 1977.

Lorenc, A. C. and Hammon, O.: Objective quality control of observations using Bayesian methods. Theory, and a practical implementation, Q. J. Roy. Meteor. Soc., 114, 515–543, doi:10.1002/qj.49711448012, 1988.

Mielke Jr., P. W., Williams, S. J., and Wu, S. C.: Covariance analysis techniques based on bivariate log-Normal distribution with weather modification applications, J. Appl. Meteorol., 16, 183–187, doi:10.1175/1520-0450(1977)016<0183:CATBOB>2.0.CO;2, 1977.

Miles, N. L., Verlinde, J., and Clothiaux, E. E.: Cloud droplet size distribution in low-level stratisform clouds, J. Atmos. Sci., 57, 295–311, doi:10.1175/1520-0469(2000)057<0295:CDSDIL>2.0.CO;2, 2000.

O'Neill, N., Ignatov, A., Holben, B., and Eck, T.: The lognormal distribution as a reference for reporting aerosol optical depth statistics: emperical tests using multi-year, multi-site AERONET sunphotometer data, Geophys. Res. Lett., 27, 3333–3336, doi:10.1029/2000GL011581, 2000.

Parrish, D. F. and Derber, J. C.: The National Meteorological Center's spectral statistical-interpolation analysis system, Mon. Weather Rev., 120, 1747–1763, doi:10.1175/1520-0493(1992)120,1747:TNMCS S.2.0.CO;2, 1992.

Perron, M. and Sura, P.: Climatology of non-Gaussian atmospheric statistics, J. Climate, 26, 1063–1083, doi:10.1175/JCLI-D-11-00504.1, 2013.

Polavarapu, S., Ren, S., Rochon, Y., Sankey, D., Ek, N., Koshyk, J., and Tarasick, D.: Data assimilation with the Candian Middle Atmosphere Model, Atmos. Ocean, 43, 77–100, doi:10.3137/ao.430105, 2005.

Rabier, F., Jarvinen, H., Klinker, E., Mahouf, J. F., and Simmons, A.: The ECMWF implementation of four dimensional variational assimilation. Part I: Experimantal results with simplified physics, Q. J. Roy. Meteor. Soc., 126A, 1143–1170, doi:10.1002/qj.49712656415, 2000.

Rawlins, F., Ballard, S. P., Bovis, K. J., Clayton, A. M., Li, D., Inverarity, G. W., Lorenc, A. C., and Payne, T. J.: The Met Office global four-dimensional variational data assimilation scheme, Q. J. Roy. Meteor. Soc., 133, 347–362, doi:10.1002/qj.32, 2007.

Razali, N. M. and Wah, Y. B.: Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, and Anderson–Darling tests, J. Stat. Model. Analytics, 2, 21–33, 2011.

Rosmond, T. and Xu, L.: Development of NAVDAS-AR: non-linear formulation and outer loop test, Tellus A, 58, 45–58, doi:10.1111/j.1600-0870.2006.00148.x, 2006.

Sauvageot, H.: The probability density function of rain rate and the estimation of rainfall by area integrals, J. Appl. Meteorol., 33, 1255–1262, doi:10.1175/1520-0450(1994)033<1255:TPDFOR>2.0.CO;2, 1994.

Sengupta, M., Clothiaux, E. E., and Ackerman, T. P.: Climatology of warm boundary layer clouds at the ARM SCP site and their comparison to models, J. Climate, 17, 4760–4782, doi:10.1175/JCLI-3231.1, 2004.

Song, H., Edwards, C. A., Moore, A. M., and Fiechter, J.: Incremental four-dimensional variational data assimilation of positive-definite oceanic variables using a logarithm transformation, Ocean Model., 54, 1–17, doi:10.1016/j.ocemod.2012.06.001, 2012.

Stephens, G. L., Vane, D. G., Boain, R. J., Mace, G. G., Sassen, K., Wang, Z., Illingworth, A. J., O'Connor, E. J., Rossow, W. B., Durden, S. L., Miller, S. D., Austin, R. T., Benedetti, A., and Mitrescu, C.: The CLOUDSAT mission and the A-train, B. Am. Meteorol. Soc., 83, 1771–1190, doi:10.1175/BAMS-83-12-1771, 2002.

Toth, Z. and Szentimrey, T.: The binormal distribution: a distribution for representing asymmetrical but normal-like weather elements, J. Climate, 3, 128–137, doi:10.1175/1520-0442(1990)0032.0.CO;2, 1990.

Trewartha, G. T. and Horn, L. H.: An introduction to climate, MGraw Hill International, London, 1971.

Yang, H. and Pierrehumbert, R.: Production of dry air by isentropic mixing, J. Atmos. Sci., 5, 3437–3454, doi:10.1175/1520-0469(1994)051<3437:PODABI>2.0.CO;2, 1994.

Zhang, C. D., Mapes, B. E., and Soden, B. J.: Bimodality in tropical water vapour, Q. J. Roy. Meteor. Soc., 129, 2847–2866, doi:10.1256/qj.02.166, 2003.
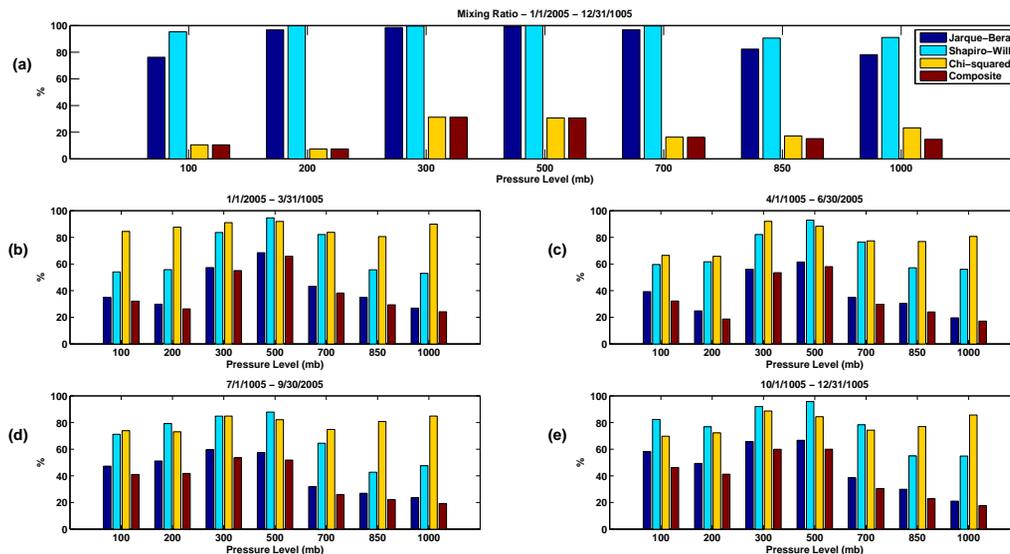
Figure 1. (a–e) Frequency of each test result on every time domain and atmospheric level. For the Jarque–Bera and Shapiro–Wilk tests, the values represent the percentage of points where the null hypothesis is rejected, concluding non-normal data. For the Chi-squared test, the frequency is the percentage of points where the null hypothesis is not rejected, demonstrating insufficient evidence against the data being lognormally-distributed. The composite test combines these results, indicating non-normal and lognormally-distributed data. The tests conclude large percentages of points where the data is non-normal and seasonal points where the data is lognormally-distributed.
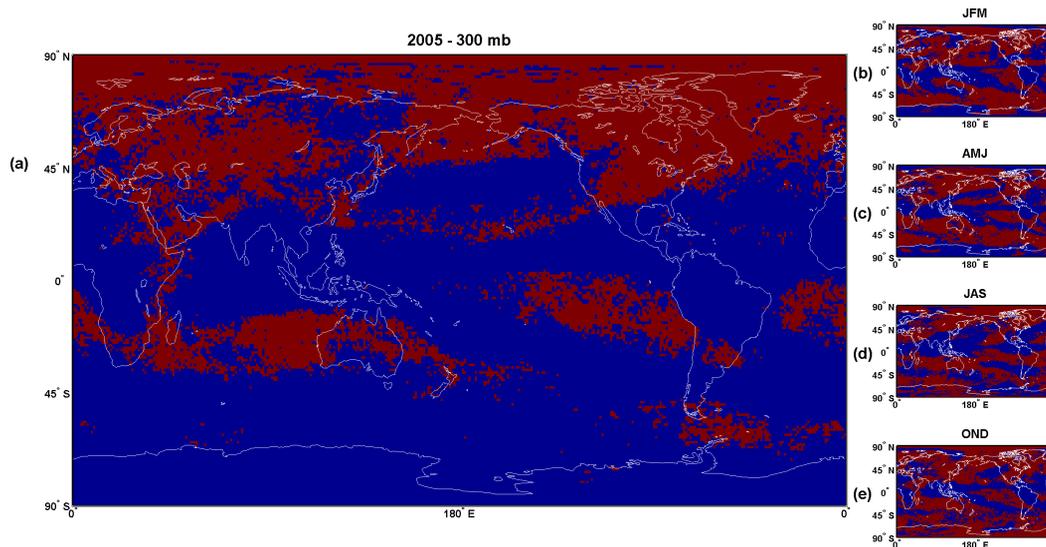
Figure 2. Composite results for water vapor mixing ratio for **(a)** 2005 and **(b–e)** each season at 300 mb.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion
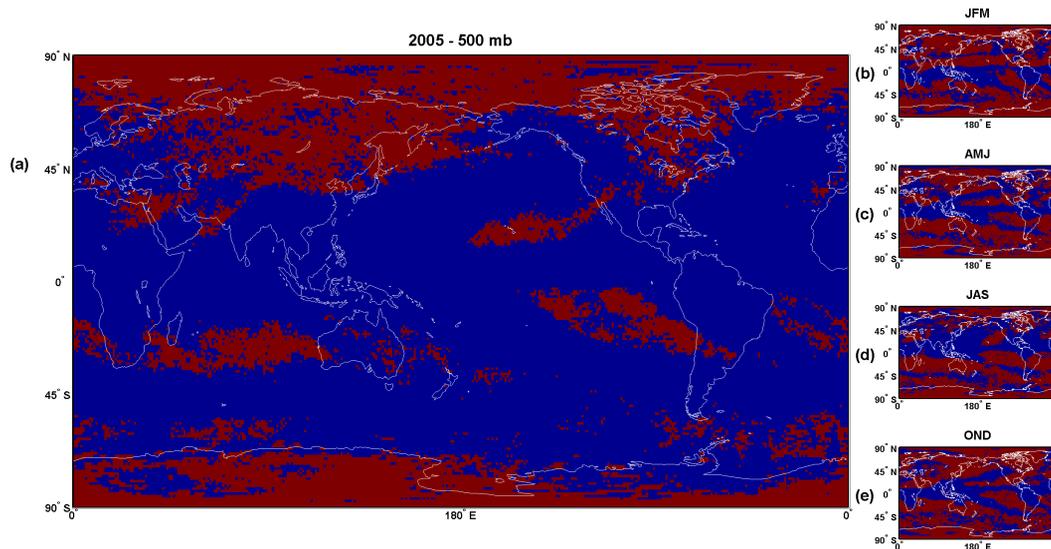
Figure 3. Similar to Fig. 2, composite results for mixing ratio for **(a)** 2005 and **(b–e)** each season at 500 mb.

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion
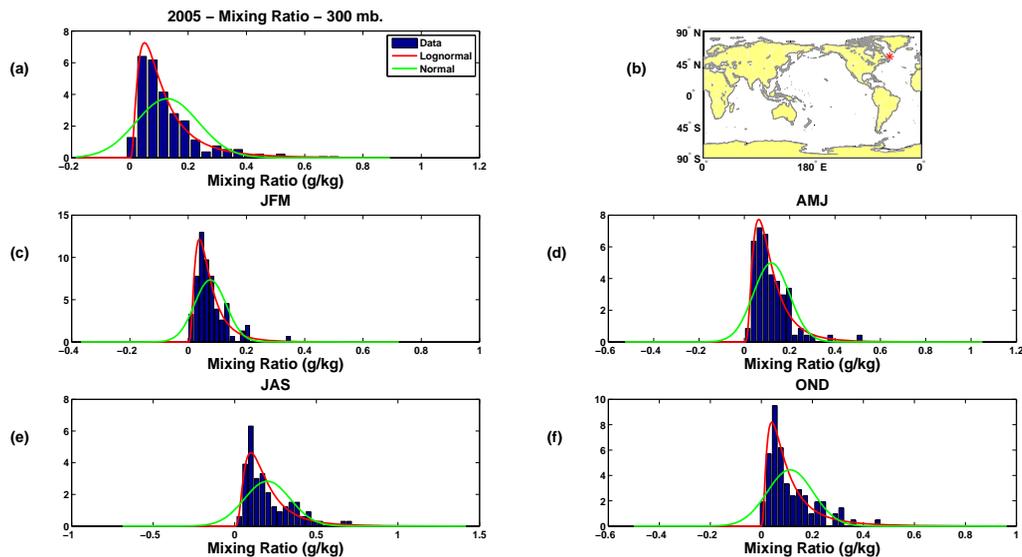
Figure 4. Histograms along with Normal and Lognormal probability distibution for **(a)** 2005 and **(c–f)** each season at 300 mb. Panel **(b)** indicates the location of this data off the Canadian eastern coast. This point is an example where each season along with the entire year of data passes the composite test indicating lognormal behavior.
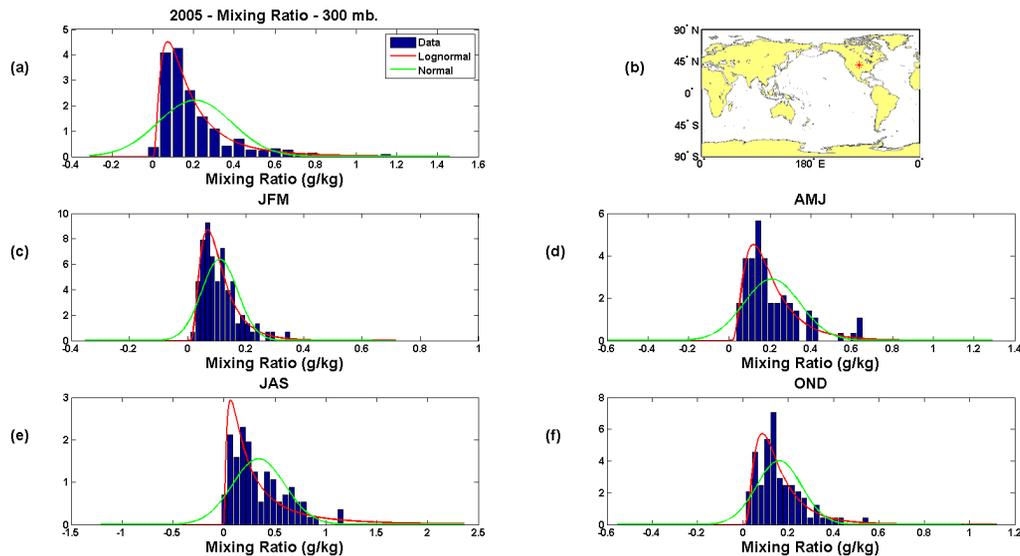
Figure 5. Similar to Fig. 4 at 300 hPa for a point in central North America. The composite test returns a positive result for 2005 and each season.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion
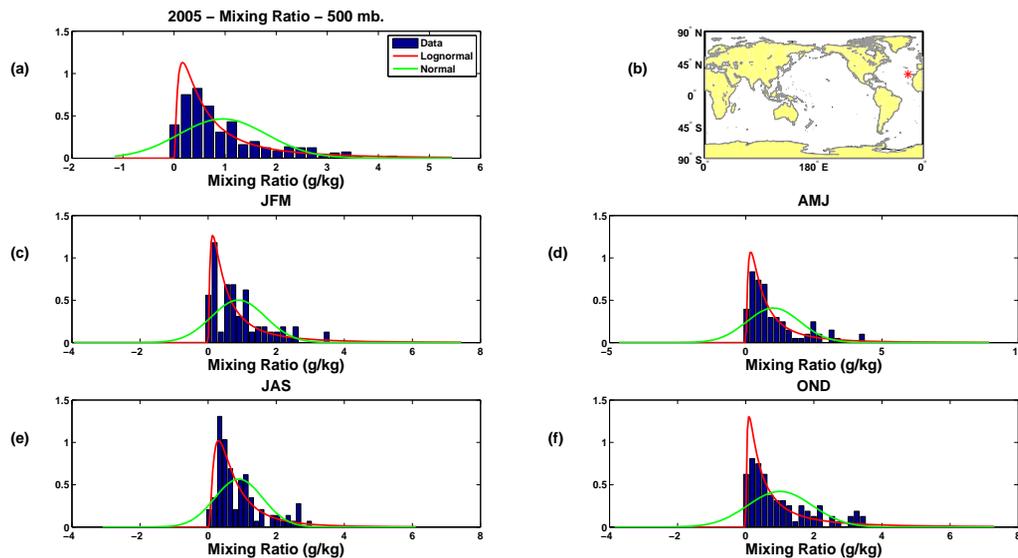
Figure 6. Similar to Fig. 4 at 500 hPa for a point in the North Atlantic. This is an example where each season, but not the entire year, passes the composite test indicating lognormal behavior.
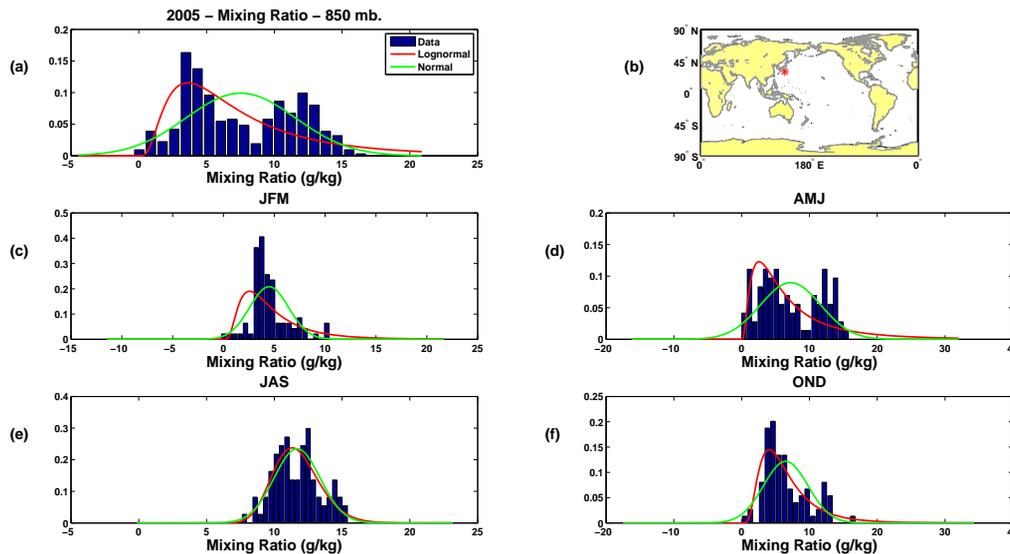
Figure 7. Location near Japan at 850 hPa where the composite test fails for every time domain.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion
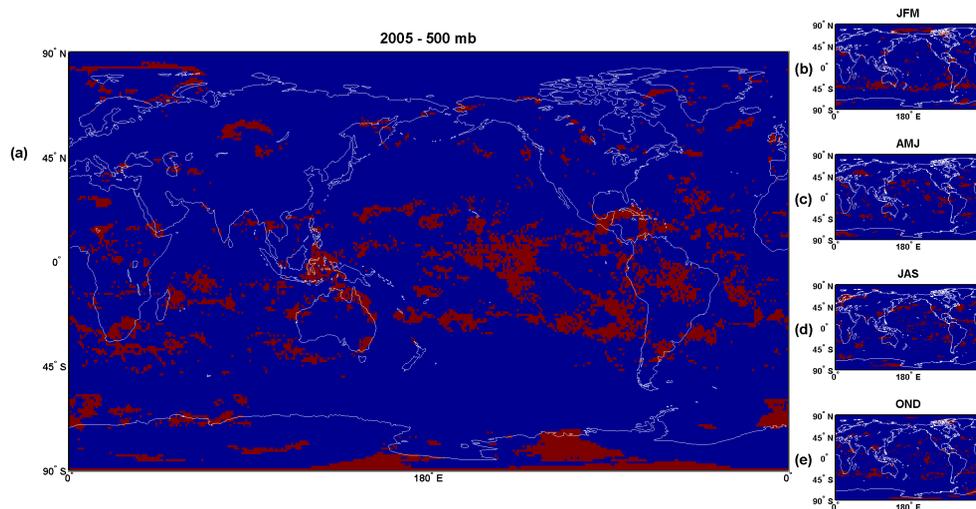
Figure 8. Similar to Fig. 2, composite results for **(a)** 2005 and **(b–e)** each season at 500 hPa for temperature.
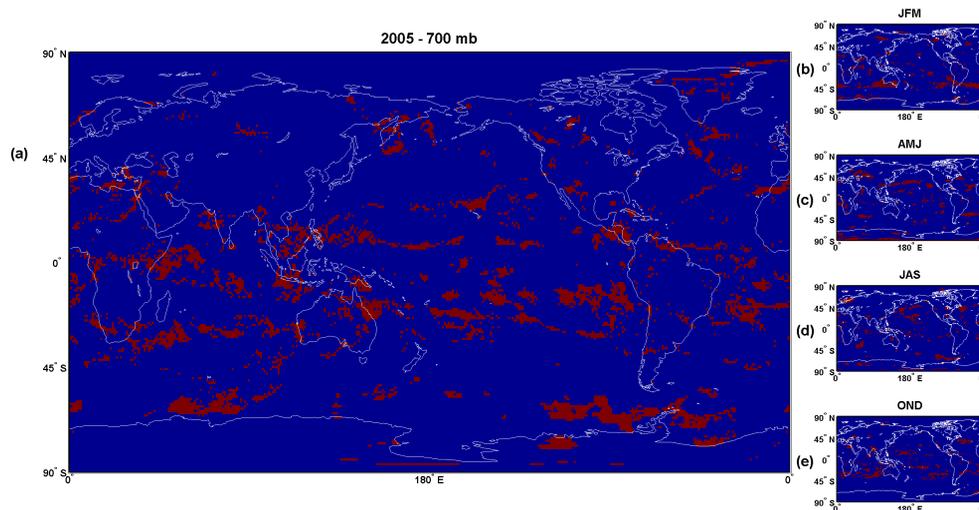
Figure 9. Similar to Fig. 2, composite results for **(a)** 2005 and **(b–e)** each season at 700 hPa for temperature.
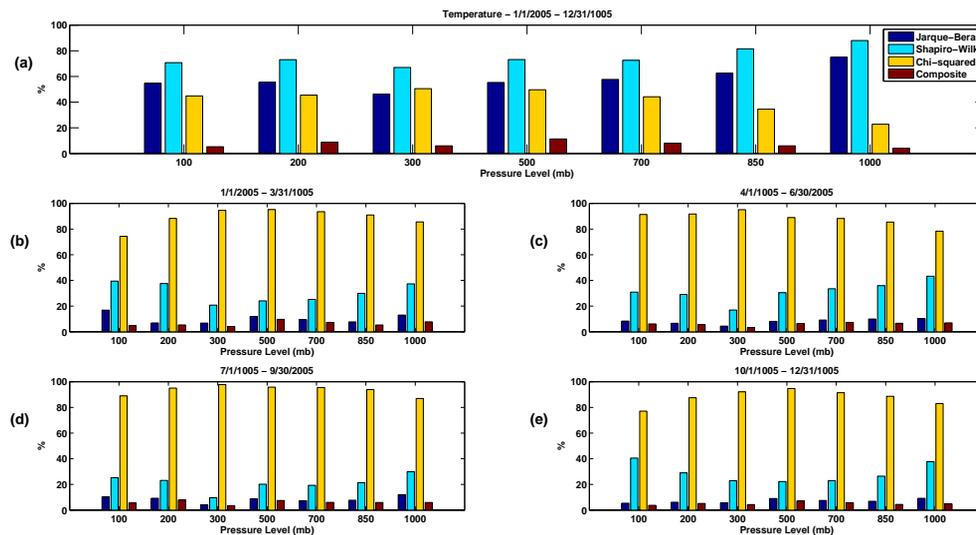
Figure 10. Frequency of each test result for temperature on every time domain and atmospheric level similar to Fig. 1. There are a significant number of points where non-normal and lognormally-distributed data appear, both annually and seasonally.
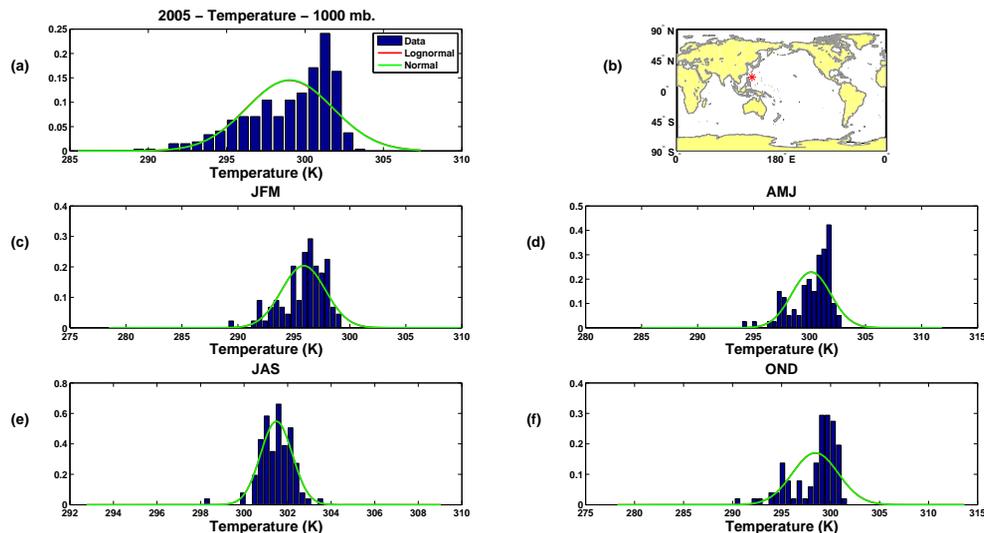
Figure 11. Temperature data for a point near Taiwan where the Shapiro–Wilk and Jarque–Bera conclude non-normally distributed data.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion
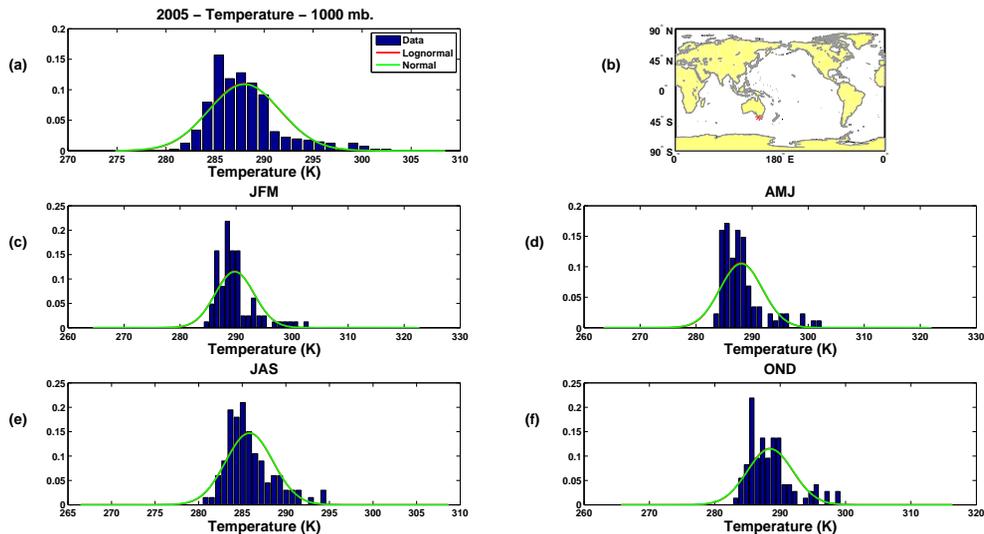
Figure 12. Temperature data for a point in Australia where the Shapiro–Wilk and Jarque–Bera conclude non-normally distributed data.
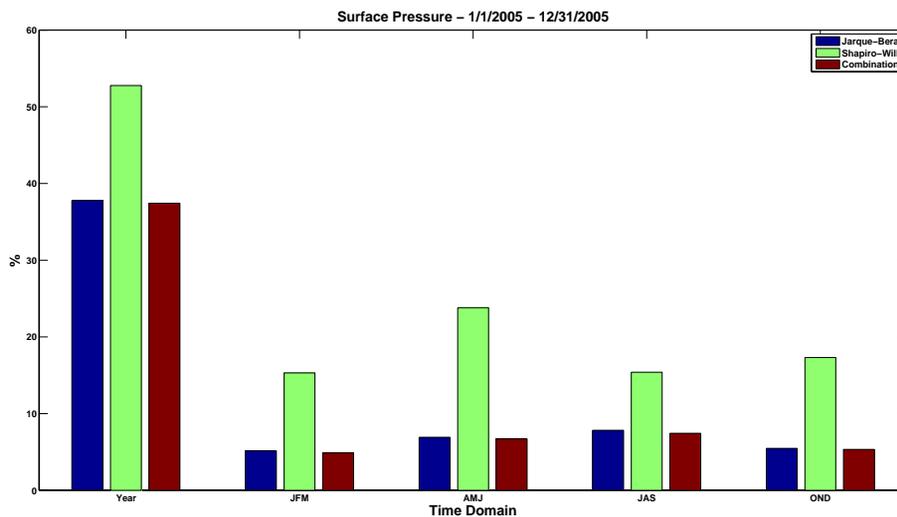
Figure 13. Similar to Fig. 1, the frequencies represent how often the normality assumption was rejected for each time domain for surface pressure.

Figure 14. (a, b) Similar to Fig. 2, the red areas indicate the normality assumption was rejected for surface pressure.

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Figure 15. Frequency (0–4) of seasons determined to be non-normal by the composite test.

Figure 16. (a–e) Similar to Fig. 1, the frequencies represent how often the normality assumption was rejected for each vertical level and time domain for the *u* component of wind.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion
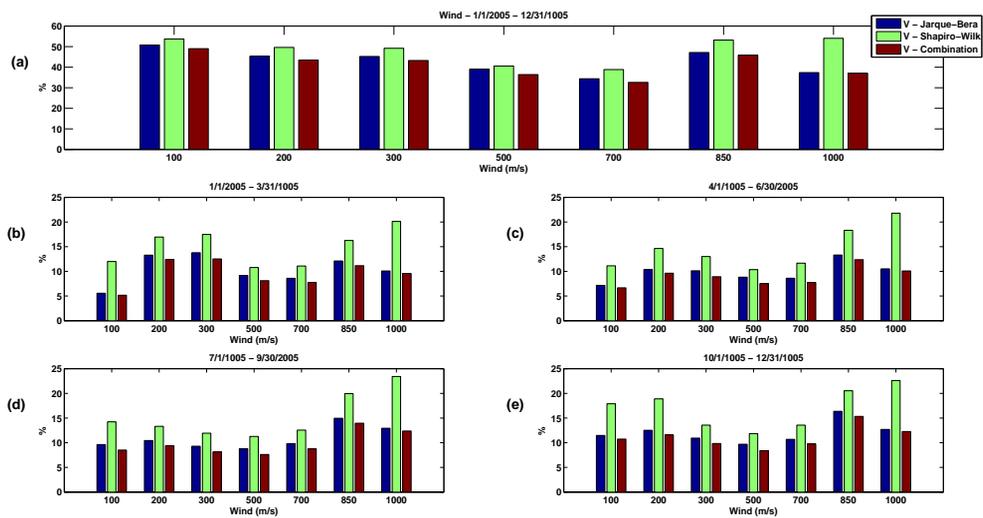
Figure 17. (a–e) Similar to Fig. 1, the frequencies represent how often the normality assumption was rejected for each vertical level and time domain for the *v* component of wind.

Figure 18. (a, b) Similar to Fig. 2, the red areas indicate the normality assumption was rejected for the *u* and *v* components of wind respectively.
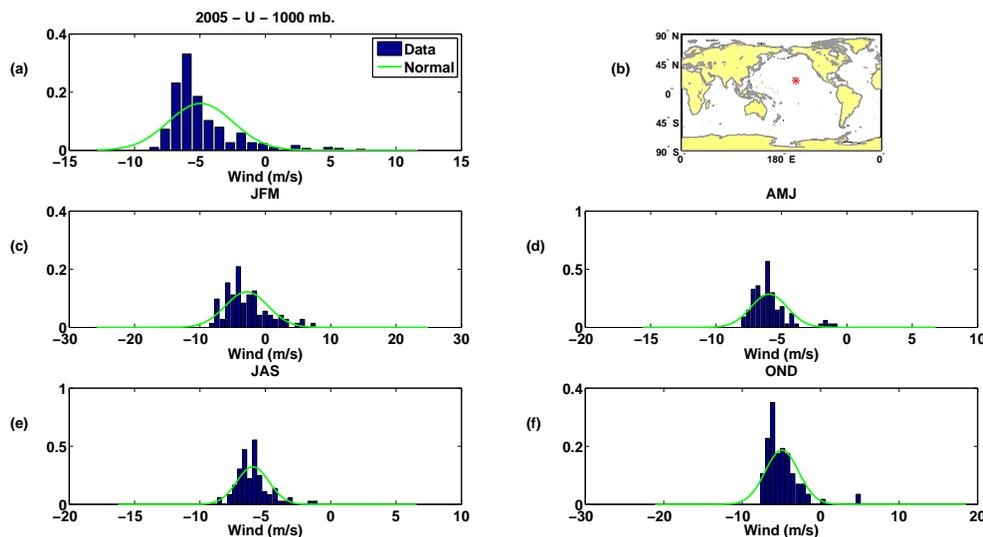
Figure 19. Similar to Fig. 4, histograms along with a normal probability distibution for **(a)** 2005 and **(c–f)** each season at 850 hPa. Panel **(b)** indicates the location of this data in the Pacific Ocean near Hawaii. In each time domain, the composite test rejected the normal assumption for this location.