Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Computational Physics 206 (2005) 334–362

JOURNAL OF COMPUTATIONAL PHYSICS

www.elsevier.com/locate/jcp

Quantifying predictability through information theory: small sample estimation in a non-Gaussian framework

Kyle Haven, Andrew Majda, Rafail Abramov *

Department of Mathematics and Center for Atmosphere/Ocean Science, Courant Institute of Mathematical Sciences, 251 Mercer Street, New York University, New York, NY 10012, USA

Received 10 September 2004; received in revised form 10 December 2004; accepted 15 December 2004
Available online 25 January 2005

Abstract

Many situations in complex systems require quantitative estimates of the lack of information in one probability distribution relative to another. In short term climate and weather prediction, examples of these issues might involve the lack of information in the historical climate record compared with an ensemble prediction, or the lack of information in a particular Gaussian ensemble prediction strategy involving the first and second moments compared with the non-Gaussian ensemble itself. The relative entropy is a natural way to quantify the predictive utility in this information, and recently a systematic computationally feasible hierarchical framework has been developed. In practical systems with many degrees of freedom, computational overhead limits ensemble predictions to relatively small sample sizes. Here the notion of predictive utility, in a relative entropy framework, is extended to small random samples by the definition of a sample utility, a measure of the unlikeliness that a random sample was produced by a given prediction strategy. The sample utility is the minimum predictability, with a statistical level of confidence, which is implied by the data. Two practical algorithms for measuring such a sample utility are developed here. The first technique is based on the statistical method of null-hypothesis testing, while the second is based upon a central limit theorem for the relative entropy of moment-based probability densities. These techniques are tested on known probability densities with parameterized bimodality and skewness, and then applied to the Lorenz '96 model, a recently developed "toy" climate model with chaotic dynamics mimicking the atmosphere. The results show a detection of non-Gaussian tendencies of prediction densities at small ensemble sizes with between 50 and 100 members, with a 95% confidence level.

© 2005 Elsevier Inc. All rights reserved.

[*] Corresponding author.
E-mail address: abramov@cims.nyu.edu (R. Abramov).

0021-9991/$ - see front matter © 2005 Elsevier Inc. All rights reserved.

doi:10.1016/j.jcp.2004.12.008

1. Introduction

Complex systems with many spatial degrees of freedom arise in environmental science in diverse contexts such as atmosphere/ocean general circulation models (GCMs) for climate or weather prediction, pollution models, and models for the spread of hazardous biological, chemical, or nuclear plumes, as well as biological molecular dynamics, complex microfluids, etc. These nonlinear models are intrinsically chaotic over many time scales with sensitive dependence on initial conditions. Given both the uncertainty in a deterministic initial condition as well as the intrinsic chaos in solutions of such systems, it is natural instead to consider an ensemble of initial data representing uncertainty in measurements and characterized by a probability density. Monitoring the propagation of such a forecast ensemble in time gives one the potential to quantify the uncertainty and measure the confidence interval and average predictive power of a single deterministic solution, whose initial condition is randomly drawn from the initial spread. Apparently, small ensemble spread at a certain time is a strong evidence of dynamics with good predictive utility, and large spread denotes otherwise. However, there are at least two major problems with ensemble simulations which are often encountered in large complex systems. The first problem arises from the ensemble prediction strategy itself: even though a qualitative estimate of “small” and “large” ensemble spreads might be enough to give some basic insight into the nature of predictability, how one can quantify the predictive utility of a forecast ensemble in a rigorous manner? The conventional way is to measure the mean and variance of an ensemble, which is equivalent to approximating the internal structure of an ensemble by a Gaussian probability density. Central issues of practical importance in an ensemble prediction such as bimodality or skewness in a forecast ensemble require a general non-Gaussian description of predictive utility. The second problem becomes important for complex systems: certainly, a larger ensemble size means better quality of a prediction. However, usually ensembles of very limited size are affordable in complex systems for making real-time forecasts, largely due to enormous consumption of computational power. Thus, the natural question arises – whether or not one can trust the information provided by a forecast ensemble with relatively small size? In other words, how one can quantify the credibility of a forecast ensemble depending on its sample size? The current work systematically addresses these two problems within the framework of information theory and rigorous predictability estimates via relative entropy.

The applicability of information theory for weather or climate prediction has been studied previously by Carnevale and Holloway [1], Schneider and Griffies [2], Roulston and Smith [3], Leung and North [4]. Recently, Kleeman [5] has suggested the relative entropy as an estimate of predictive utility in an ensemble forecast relative to the climatological record, as well as a signal-dispersion decomposition. The Gaussian framework of relative entropy and signal-dispersion decomposition has been tested by Kleeman et al. [6] for a simple 100-mode truncated Burgers–Hopf model with chaotic behavior and well-understood spectrum and autocorrelation time scaling (for complete model description and climatology see Majda and Timofeyev [7,8], and Abramov et al. [9]). Majda et al. [10] developed a more sophisticated framework of predictability through relative entropy for non-Gaussian probability density functions, which includes a hierarchy of rigorous lower bounds on relative entropy through the statistical moments beyond the mean and covariance through maximum entropy optimization (Mead and Papanicolaou [11]). Abramov and Majda [12] converted the non-Gaussian predictability framework into a practical tool through the hierarchy of lower bounds and a rapid numerical optimization algorithm. Recently, Cai et al. [13] exhaustively tested several facets of the non-Gaussian information theoretic predictability framework in a simple chaotic mapping model with an explicit attractor ranging from Gaussian to fractal as parameters are varied. Kleeman and Majda [14] have quantified the loss of information in coarse-grained ensemble estimators and applied these ideas to geophysical turbulence. Different applications of relative entropy as a predictability tool were developed in Abramov and Majda [12]; besides a straightforward measure of lack of information in the climate relative to the prediction ensemble, the relative entropy can be used in estimating lack of information in a forecast ensemble relative to the actual events (Roulston and Smith [3]), in evaluating additional information content in the skewness and higher order forecast moments (non-Gaussianity), and the information flow between different subsets of phase space in an ensemble forecast (statistical correlation between different large scale phenomena). In Abramov and Majda [12] all of these facets were demonstrated for the Lorenz '96 model (Lorenz and Emanuel [15]), including highly non-Gaussian behavior. Finally, Abramov et al. [16] successfully applied the relative entropy framework to the simplest midlatitude atmospheric climate model, barotropic T21 spherical truncation with realistic orography in two different dynamical regimes, with each regime mimicking the behavior of atmosphere at a certain height. In particular, the information flow was found responsible for correlated switches in large scale structures like the Arctic Oscillation, North Atlantic Oscillation, and Pacific/North American pattern. All of the above work demonstrates many practical facets of quantifying uncertainty in ensemble forecasts through the relative entropy; however, all of the work in idealized settings described above utilized large ensemble sizes. This points toward the central issue of quantifying the uncertainty of an ensemble forecast with limited ensemble size in the non-Gaussian framework of information theory and relative entropy. This is the main topic of this paper.

Perhaps, the most sophisticated contemporary uses of ensemble predictions in complex systems with small ensemble size occur in weather and climate predictions. Studies of predictions with forecast ensembles were performed, among others, by Anderson and Stern [17], Toth and Kalnay [18] and Palmer [19]. Practice shows that with current computational facilities, in order to perform real-time forecasts with large atmospheric GCMs, the size of the prediction ensemble has to be small, about 10–50 members, depending on spatial resolution, with the possibility of 50–100 member ensembles in the near future. Due to limited forecast ensemble size, certain complications arise concerning the credibility of information provided by such ensemble. The common strategy of dealing with small forecast ensembles is to maximize the information provided by the limited sample size, via generating a forecast ensemble in a very specific way. In particular, Ehrendorfer and Tribbia [20] show that for correct error growth reconstruction, the fastest growing directions of the phase space have to be sampled. Two efficient methods of ensemble generation are usually used in practical ensemble forecasts: local Lyapunov vectors (Toth and Kalnay [18], Kalnay [21]) and singular vectors (Palmer et al. [22], Reynolds and Palmer [23]). The efficiency of prediction depending on ensemble size has been studied previously by Buizza and Palmer [24].

A novel strategy, which may successfully complement the existing one described above, is developed here within an information theory predictability framework. Rather than providing the way of generating a forecast ensemble in a most efficient manner, this strategy evaluates the credibility of information in an existing ensemble (however generated), and evaluates its information content in a rigorous manner through an appropriate modification of the relative entropy in a forecast. Two general techniques are devised here to account for the lack of information due to small sample size: one is based on the statistical method of null-hypothesis testing, while the other employs a central limit theorem for the relative entropy of non-Gaussian moment-based probability densities. These two methodologies are systematically compared against the straightforward "perfect predictability" method, i.e., when measured information is assumed to be precise regardless of the sample size. The techniques are tested in both the "lab" and "field" set-ups: in the "lab" set-up the methodologies are utilized for relative entropy of an explicitly defined, but statistically undersampled, family of probability density functions with parameterized skewness and bimodality; in the "field" set-up the framework is tested for the Lorenz '96 system, which is a simple forty-dimensional model with chaotic behavior and unstable wave structure like that in a realistic complex weather system (Lorenz [25], Lorenz and Emanuel [15], Abramov and Majda [12]).

Section 2 begins the technical discussion of relative entropy and ends with an outline of the remainder of the paper.

2. Measuring predictability through relative entropy

As is well known in information theory [26], the relative entropy

$$R(p|q) = \int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x \tag{1}$$

is a measure of the average lack of information in one probability density function (PDF), q, relative to some other PDF, p. Formally, it can also be thought of as the information in p relative to q. The relative entropy is a nonsymmetric, convex functional in p with the property that $R(p|q) \geqslant 0$ with equality if and only if $p = q$. It can therefore be thought of as a nonsymmetric distance between p and q. Here, the relative entropy is used as an indicator of predictability, or predictive utility, in that it measures the utility of a particular prediction strategy q as compared to an unknown prediction strategy p, where p depends upon perfect knowledge of the underlying system. The relative entropy is an attractive measure of predictability for many reasons including that it is invariant under arbitrary changes of variables [27] and that, for a general class of densities, it can be decomposed into its signal and dispersion components [10].

The physical interpretation of p and q depends upon the particular setting. For instance, for long term climate prediction, p may represent the time dependent prediction PDF that can theoretically be found by solving the Liouville equation associated with the original dynamics. In this scenario, q represents the equilibrium PDF or climate, and $R(p|q)$ quantifies the amount of information that p provides beyond q. In weather and short term climate prediction, prediction strategies usually only utilize the first and second moments [21,18]. Here again, p may represent the perfect prediction PDF theoretically derived from the dynamics, and q a Gaussian PDF representing the prediction strategy using the first two moments. In this case, $R(p|q)$ measures the effectiveness of the two-moment strategy as compared to the perfect prediction scenario.

One of the main obstacles in straightforward use of the relative entropy as a measure of predictability is that large number of degrees of freedom for p and q make standard integration techniques impractical. When both p and q are known to be Gaussian, as is well known, the calculation of the relative entropy simplifies to an algebraic expression in terms of the moments of the two densities [5,6]. In [10], a hierarchical procedure for obtaining a lower bound estimate of the relative entropy based solely on the moments of p and q is described. Under the assumption that q is of a certain form, which includes Gaussian densities as a special case, this moment-based relative entropy estimate can be computed in a straightforward manner even in large-dimensional spaces by a sum of one-dimensional and two-dimensional entropies [12,16]. For ease of reading, the term entropy moment (EM) estimate is used in place of the more cumbersome moment-based relative entropy estimate. Applications of the methodology to ensemble predictions were already noted in the introduction [12,13,16].

2.1. Small sample variability

In the moment-based studies of predictability mentioned earlier, large ensemble sizes were used and it was always assumed that the moments of p and q are known precisely. In practice, it often happens that only a finite sample from p is given. Additionally, the high computational cost of generating the data can lead to small sample sizes as compared to the large degrees of freedom of the system. Thus, only imperfect estimates of the moments of p are known. The term perfect predictability is used throughout the paper when referring to the EM estimate with perfect knowledge of the moments of p. Due to the complicated form of the EM estimate, it is not clear how to determine an unbiased estimate of this quantity based on the sample moments. Instead, the methodology of the EM estimate is usually carried out with the sample moments replacing the actual moments. This will be referred to as the sample EM estimate. The variability of the sample moments will lead to variability in the sample EM estimate. If the sample size is large, then the sample EM estimate will be roughly equal to the EM estimate. However, for small sample sizes, the sample EM estimate may be much more or much less than the EM estimate, leaving no clear conclusion about the level of predictability for the prediction strategy. The main focus of this paper is the general behavior of the sample EM estimates and the introduction of a minimum predictability, with a statistical level of confidence, which is implied by the data. The latter requires a shift from trying to estimate the average lack of information in q relative to p using a finite sample from p to directly quantifying the lack of information in q relative to the sample itself.

2.2. Sample utility

Formally, if $R(p|q)$ represents the information content of p beyond q, then having only a finite sample from p should necessarily decrease the information content. This can also be thought of as a loss of information due to sample estimation. As an example, consider the case when p is very close to q, so that $R(p|q)$ is small. Since p is different from q, the positive value of the relative entropy indicates that there is information content in p. However, if a small random sample is chosen using p, then it will be difficult to statistically show that p is different from q to any reasonable level of confidence. That is, that the random sample could not have possibly come from q. Thus, the information content of the random sample should be zero. As the sample size increases, it is reasonable to expect that the information content of the random sample will statistically increase and approach $R(p|q)$.

In this paper, any formal estimate of the random sample's information content is referred to as a *sample utility*. The sample utility can also be thought of as a measure of the unlikeliness that the random sample from p came from q. The properties that a sample utility should possess, which were discussed in the previous paragraph, are listed in the following definition.

Definition 1. The sample utility should possess the following three properties:

- *The sample utility should be less than the relative entropy.* A random sample provides less information than having p itself. This can also be thought of as a loss of information due to sample estimation. With the same reasoning, a sample utility which is based only on the sample moments should be less than the EM estimate.
- *The sample utility should statistically increase with sample size.* Larger sample sizes provide more information.
- *The sample utility should approach the relative entropy as the sample size approaches infinity.*

Since extremely uncharacteristic random samples can be theoretically produced by q, it is impossible to define a positive estimate for a sample utility which will remain below the relative entropy 100% of the time. Confidence levels are introduced to allow for the statistically rare occurrences of these uncharacteristic random samples. In relation to the actual predictability, then, the sample utility should be the minimum amount of predictive utility, with a statistical level of confidence, which is implied by the data. The level of confidence is arbitrary and chosen to be 95% here for demonstration purposes.

Any statistical method which is used to test whether a random sample comes from a particular distribution, q, can be used to test for zero versus nonzero relative entropy. Thus, distinguishing zero versus nonzero sample utilities is not a difficult task. However, quantifying the sample utility when it is not zero requires more work.

In order to define a measure of sample utility that tends to be less than and statistically increasing to the true relative entropy, it is necessary to consider all the densities that could have reasonably produced the data and choose the minimum relative entropy over this group. Thus, the method of defining the group of admissible densities should theoretically determine the measurement of the sample utility. In this paper, two strategies are suggested based on standard statistical methods. The first strategy, which is outlined in Section 4, incorporates a confidence ellipsoid for the moments of q. Whether or not the sample moments of p fall within the confidence set determines whether there is any sample utility. That the confidence ellipsoid shrinks to a point as the sample size approaches infinity implies the last two properties in Definition 1. This measure of sample utility is very conservative and has little trouble satisfying the first property. The second strategy, outlined in Section 5, deals directly with a one-sided confidence interval for the EM estimate. Theoretically, the lower bound of this confidence interval will be below the EM estimate 95% of the time and will statistically increase as the sample size approaches infinity. The confidence interval depends on a central limit theorem which is stated and proved in Appendix B.

An important issue in measuring the sample utility is the number of moments that should be incorporated into the estimate. It is sometimes argued that often only the first two moments are adequately approximated by the sample moments and thus the higher moments should be neglected. In reality, the importance of the higher moments is not only dependent upon the sample size, but also on the difference between the moments being compared. Ideally, a measure of sample utility should automatically compensate for the greater variability of the higher moments, so that no decision about whether to include higher moments is needed.

In the following section, the perfect predictability methodology for the EM estimates is briefly outlined. A discussion of the behavior of the sample EM estimates follows.

3. Perfect predictability methodology

In a perfect predictability scenario, it is often assumed that the prediction PDF, p, is known precisely. Here, the term perfect predictability is used to refer to the perfect knowledge of the first K moments, which are denoted by $\mathbf{M}(p) = (M_1(p),\ldots,M_K(p))$. For this paper, the value of K is taken to be 2 or 4.

When $\mathbf{M}(p)$ is known precisely and q is of a certain exponential form (see Eq. (A.2)), then a lower bound estimate of $R(p|q)$ is found by minimizing $R(\rho|q)$ over all probability densities satisfying $\mathbf{M}(\rho) = \mathbf{M}(p)$. The convexity of R ensures that the minimum will be reached for some PDF, which shall be refer to as the entropy moment (EM) PDF [10]. The lower bound estimate, which shall be refer to as the EM estimate, can be computed numerically by standard optimization procedures for small values of K and small degrees of freedom [12]. If $K = 2$, then both q and the EM PDF are Gaussian densities.

The EM estimate of $R(p|q)$ may also be thought of as the exact value of the minimum information content given only $\mathbf{M}(p)$. More precisely, let $\boldsymbol{\alpha}$ denote an admissible set of moments. Then there is an infinite family of densities that can produce these moments. The minimum information content over the family is given by

$$P(\boldsymbol{\alpha}|\mathbf{M}(q)) = \min\{R(\rho|q) : \mathbf{M}(\rho) = \boldsymbol{\alpha}\}. \tag{2}$$

There is no maximum information content over the family. In this new notation, the EM estimate is written $P(\mathbf{M}(p)|\mathbf{M}(q))$. The notation emphasizes the fact that the EM estimates of relative entropy depend solely on the moments of p and q. Since P is convex in $\boldsymbol{\alpha}$ with the property that $P(\mathbf{M}(p)|\mathbf{M}(q)) \geqslant 0$ with equality if and only if $\mathbf{M}(p) = \mathbf{M}(q)$, it can be thought of as a nonsymmetric distance between $\mathbf{M}(p)$ and $\mathbf{M}(q)$.

When only a finite sample estimate of $\mathbf{M}(p)$ is given, the EM methodology may still be carried out with the sample moments in place of the true moments. This will be referred to as the sample EM estimate. This is the strategy implemented in earlier work using these ideas for ensemble prediction for large ensemble sizes [12,13]. Let $\mathbf{s}$ denote the sample estimates of $\mathbf{M}(p)$. Then $P(\mathbf{s}|\mathbf{M}(q))$ represents the sample EM estimate. As previously discussed, for small sample sizes, $P(\mathbf{s}|\mathbf{M}(q))$ can be much more or much less than $P(\mathbf{M}(p)|\mathbf{M}(q))$. Often, the sample EM estimate is higher than and decreases to the EM estimate as the sample size increases through moderate values. This tendency is exactly the opposite of the desired properties listed in Definition 1. This behavior can be observed in Figs. 2, 3, 5 and 6, where the performance of the three methodologies, described in this paper, are compared for specific densities p and q. Throughout, q is taken to be the standard Gaussian density, while p is the EM PDF with moments given in the table of the first panel. The second panel shows the graphs of p and q. The third panel is a series of boxplots which show the results from using the EM methodology for various sample sizes. For each sample size, 500 independent ensembles are generated (thus constituting a super-ensemble), which are used to compute $P(\mathbf{s}|\mathbf{M}(q))$. Each boxplot shows the range of the results for the 500 ensembles with the box representing the middle 50%, and the horizontal line through the box represents the median value. The horizontal dashed line across all the boxplots represents the EM estimate $P(\mathbf{M}(p)|\mathbf{M}(q))$. In Figs. 2 and 5, only two-moment estimates are shown. In Figs. 3 and 6, the corresponding four-moment estimates are shown.

The first set of figures show a case where p is relatively close to the standard Gaussian q, with $R(p|q) = 0.06016$. It is therefore not surprising to see that, for small sample sizes, the majority of the sample EM estimates lie above the EM estimate. Specifically, at sample size 25 in Fig. 3, the sample EM estimates can reach values of over 0.6. Practically speaking, this means that the EM methodology applied directly to the sample moments with small ensemble size can lead to extreme over estimation of predictability. For the two-moment case, in Fig. 2, this over estimation is even more pronounced, since the two-moment EM estimate is zero. Not surprisingly, the tendency of the sample estimate to over estimate the EM estimate becomes more pronounced as the EM estimate approaches zero. In Figs. 5 and 6, p is further from the standard Gaussian q, with $R(p|q) = 0.7971$. Here, the sample estimates, even for small sample sizes, are more evenly distributed about the EM estimate. However, the spread of the data is much wider than in the first case. The distribution of the data is actually irrelevant here. In practice, only one data point will be given. The larger variability only means that sample EM estimates from this PDF are more likely to be further from the true EM estimate. These two examples illustrate how the variability of the EM methodology applied to the sample moments can severely distort the truth about the EM estimate.

It is interesting to consider the difference between the two-moment and four-moment cases, since it is a topic of debate whether higher moments are reliable enough to include in estimates of predictability. Even for sample size 25 in Fig. 5, the majority of the sample estimates fall below the true value of 0.7971, whereas in Fig. 6, the data roughly centers about the true value. This indicates that the EM methodology can be sensitive to the higher moments, even for extremely small sample sizes, and suggests that the higher moments should not be discarded arbitrarily. However, this does not indicate that the four-moment sample EM methodology performs better than the two-moment methodology for small sample sizes. Indeed, only one random sample is typically given, and whether this sample is representative of the underlying density is impossible to know. Also, since there is more variability in the higher moments than in the lower moments, it is not surprising to note that there is more variability in the four-moment EM estimates than in the two-moment EM estimates. All this is more evidence of the need for a measure of the sample utility.

If p is an EM density, then $P(\mathbf{M}_4(p)|\mathbf{s}_2)$ measures the amount of information contained in p beyond the Gaussian density with mean and variance calculated from the random sample and is referred to here as the *non-Gaussianity*. The subscripts refer to the number of moments that are used. The quantity $P(\mathbf{s}_4|\mathbf{s}_2)$ is therefore the sample EM estimate of the non-Gaussianity. Figs. 4 and 7 show the same boxplot information as in the previous figures, but instead of computing $P(\mathbf{s}|\mathbf{M}(q))$, for each ensemble, $P(\mathbf{s}_4|\mathbf{s}_2)$ is computed. The distribution of the data for the boxplots in these figures collaborates the observations already discussed. Namely, for p close to Gaussian, sample estimates tend to over estimate more, but have smaller variability.

The general behavior of the EM methodology is captured by these two examples. For larger relative entropies, the distribution of the sample EM estimates tends to spread evenly about the EM estimate, while for smaller relative entropies, most of the sample EM estimates lie above the EM estimate. As the true relative entropy decreases, the sample EM methodology becomes increasingly biased. The span of the data increases with the number of moments being considered as well as the difference between p and q. This methodology tends to violate the first two sample utility properties listed in Definition 1. Namely, the sample EM methodology leads to estimates that can be larger than and statistically decreasing to the true EM estimate.

In the following section, a method based on hypothesis testing is developed in an attempt to define a measure of the sample utility that possesses the three desired properties in Definition 1.

4. Adjusted moments methodology

In this section, a measure of the sample utility is defined using the statistical method of hypothesis testing. Since the EM estimate is completely dependent upon the mean and centered moments, it is natural to question whether or not the sample mean and centered moments could have possibly been produced by q instead of p. Since the object is to determine the degree to which the data indicates a difference between p and q, it is assumed that the two densities are the same, prior to the observation of the data. In the context of hypothesis testing, this translates to a null hypothesis of

$$H_0 : p = q. \tag{3}$$

This initial assumption sets a bias towards a lack of predictability for the sample. Given the sample data from p, sample statistics may then be used to try to prove that the hypothesis is incorrect to some level of confidence. For demonstration purposes, the level of confidence is set to 95% throughout the paper.

As stated in Section 2, there are many tests that can be employed to determine whether a set of data may have been produced by a particular distribution q. If a statistical test is not able to prove, with 95% confidence, that the data did not come from q, then the possibility that the data came from q cannot be ruled out, and thus the sample utility must be zero. On the other hand, if the test indicates that the data probably did not come from q, then it is safe to say that the sample utility is positive. The only problem is that the magnitude of the sample utility is not specified. It may be possible to devise a method of measuring sample utility based on the p-value of a statistical test, but this idea is not explored here.

4.1. Adjusted moment algorithm

In an attempt to quantify positive sample utilities, a statistical test based on a 95% confidence set for the sample moments is employed. Let $\mathscr{E}(\mathbf{M}(q))$ denote a K dimensional set where, 95% of the time, if a sample is chosen from q, the sample moments will fall within the set. The shape of $\mathscr{E}(\mathbf{M}(q))$ is chosen to be an ellipsoid (see Eq. (A.11)) to coincide with the asymptotic confidence ellipsoid given by a central limit theorem proved in Proposition A.3. A simple test of the null hypothesis is whether the sample moments, $\mathbf{s}$, falls within $\mathscr{E}(\mathbf{M}(q))$. If $\mathbf{s}$ does not fall within $\mathscr{E}(\mathbf{M}(q))$, then the null hypothesis is rejected and thus $R(p|q) > 0$. If $\mathbf{s}$ does fall within $\mathscr{E}(\mathbf{M}(q))$, then there is not enough evidence to reject the hypothesis that $p = q$, and so the possibility that $R(p|q)$ could be zero cannot be ruled out. The algorithm is schematically shown in Fig. 1. When the moment test fails to reject the null hypothesis (denoted by **S1** in Fig. 1), the sample utility is defined to be zero. However, when the null hypothesis is rejected, only a formal attempt can be made to measure the sample utility. Note that $\mathbf{s}$ lies within the confidence ellipsoid if and only if $\mathbf{M}(q)$ falls within $\mathscr{E}(\mathbf{s})$, where $\mathscr{E}(\mathbf{s})$ denotes the confidence set $\mathscr{E}(\mathbf{M}(q))$ centered at the point $\mathbf{s}$. Thinking of P as a nonsymmetric distance between $\mathbf{s}$ and $\mathbf{M}(q)$, it is natural to define the sample utility as the minimum P-distance between $\mathscr{E}(\mathbf{s})$ and $\mathbf{M}(q)$, as demonstrated for the moment set **S2** in Fig. 1). This measure of sample utility assigns a value of zero when $\mathbf{s}$ falls within the confidence ellipsoid and, in effect, moves those values of $\mathbf{s}$

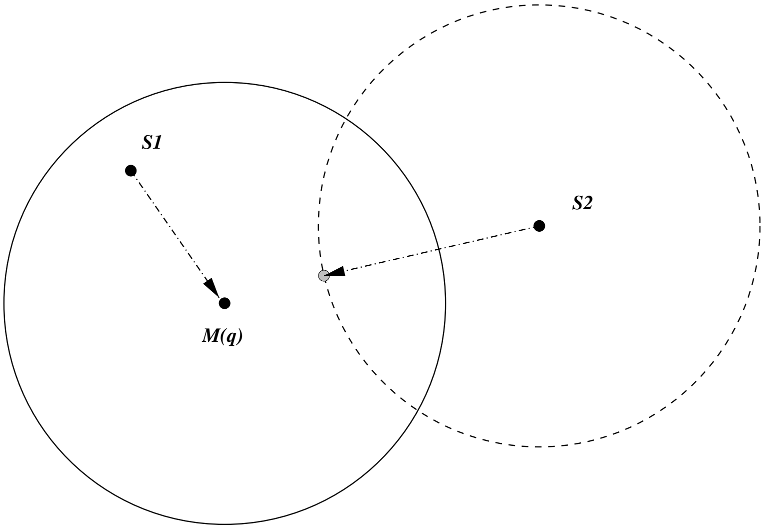


Fig. 1. *AM schematic*: Sample moment set **S**1 falls within confidence ellipsoid $\mathscr{E}(\mathbf{M}(q))$ and therefore is adjusted to $\mathbf{M}(q)$ (null-hypothesis is not rejected). Sample moment set **S**2 does not fall within $\mathscr{E}(\mathbf{M}(q))$ (null-hypothesis is rejected), however **S**2 is adjusted to a point on $\mathscr{E}(\mathbf{S}2)$ with minimal distance from $\mathbf{M}(q)$.

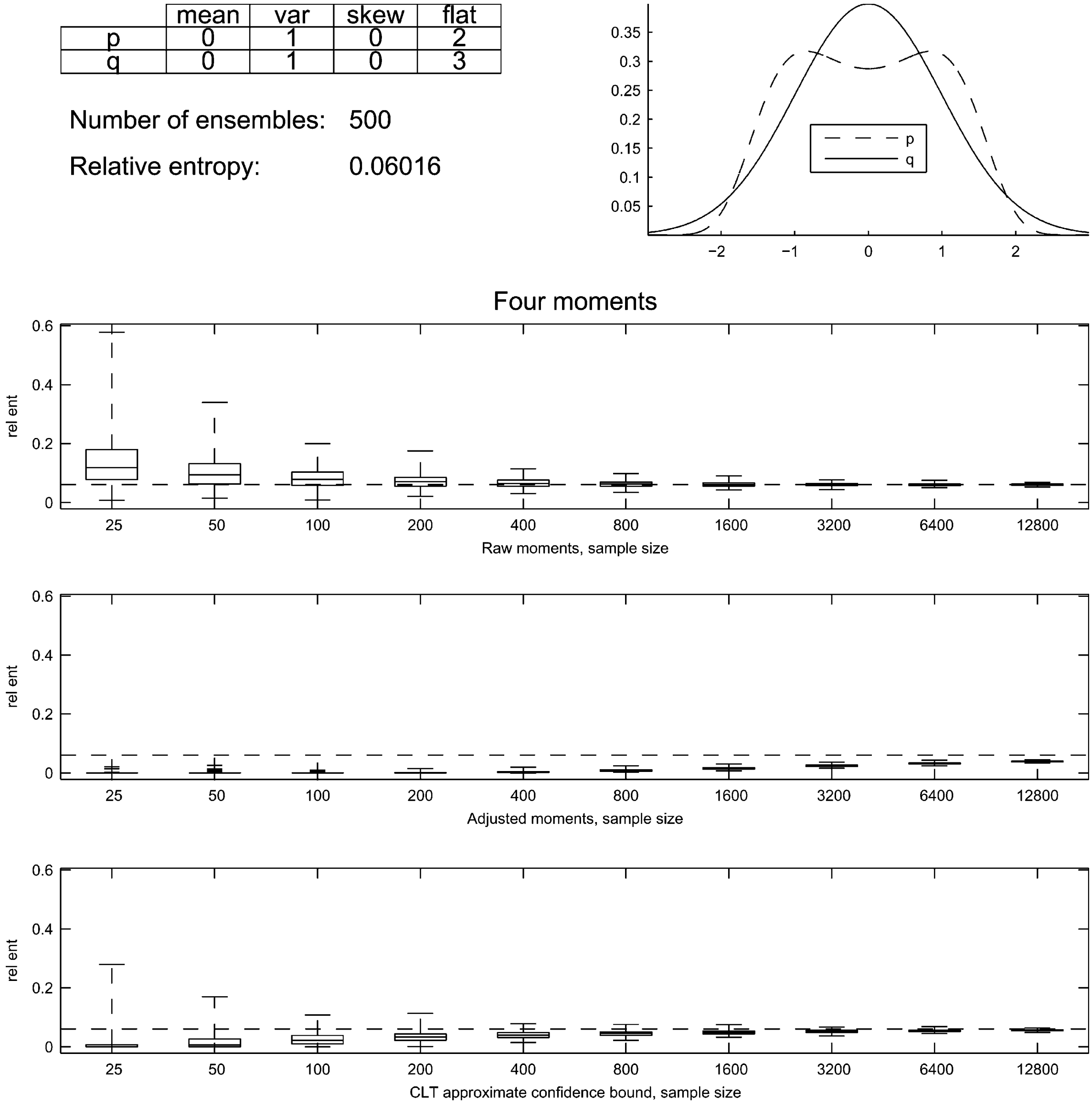
which are outside of the confidence ellipsoid closer to $\mathbf{M}(q)$. For this reason, this measure of the sample utility is called the adjusted moment (AM) sample utility.

The AM sample utility seems to possess the desired properties listed in Definition 1. First, if the true moments $\mathbf{M}(p)$ lie within $\mathscr{E}(\mathbf{s})$, then the AM sample utility will be less than the EM estimate. If p is close to q, then this should occur roughly 95% of the time. In fact, since the minimum over $\mathscr{E}(\mathbf{s})$ is used, this property is almost never violated. Second, since the confidence ellipsoid shrinks to a point as the sample size goes to infinity, it is reasonable to expect that the AM sample utility will statistically increase to the EM estimate.

In the fourth panel of Figs. 2–7, a similar series of boxplots to the third panel are shown for this adjusted moments methodology. For each figure, at least 95% of the AM sample utilities fall below and statistically increases to the EM estimate. Thus, for each random sample, the EM estimate lies above the AM sample utility, with at least 95% confidence.

In order to obtain an executable algorithm for the adjusted moment methodology, some simplifying assumption are made. In the case where only one moment is constrained, no additional assumptions need to be made. In this case, a 95% confidence interval for the moment can be computed using a bootstrap method for small sample sizes or an approximate 95% confidence interval given by the central limit theorem for large sample sizes. If the moment from q lies outside of the shifted confidence interval, then the convexity of P implies that the minimum P-distance between the moment and the shifted confidence interval occurs at the endpoint of the confidence interval that is closest to the moment of q. For the multiple constraint case, finding the point where the minimum P-distance occurs is not as easy. The convexity of P implies only that the point lies on the boundary of $\mathscr{E}(\mathbf{s})$ which is closest to $\mathbf{M}(q)$. In order to get an executable algorithm, it is assumed that the appropriate point lies on the line connecting **s** to $\mathbf{M}(q)$. At the very least, this choice of adjusted moments ensures that the resulting sample utility will be less than $P(\mathbf{s}|\mathbf{M}(q))$.

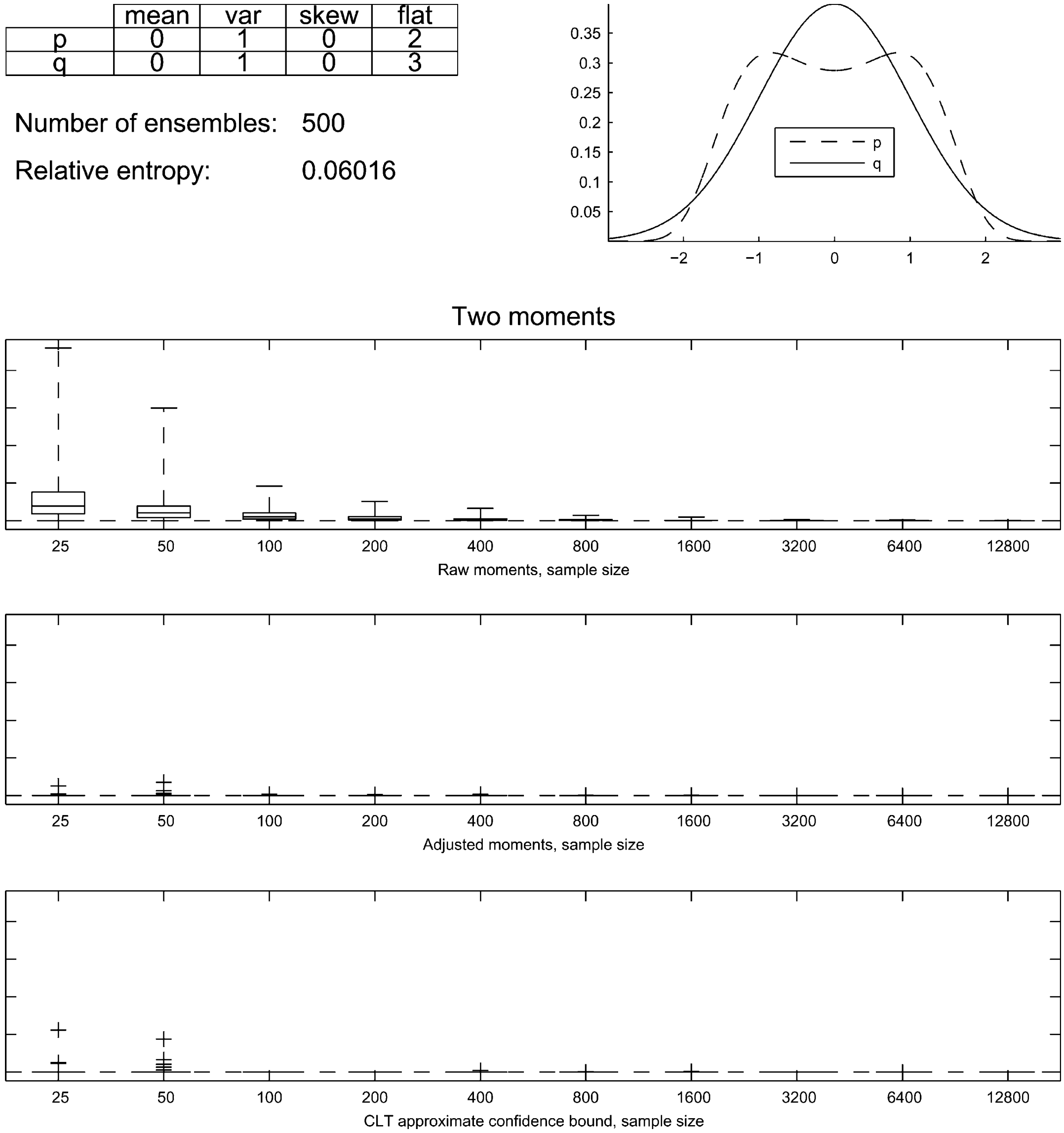
As an alternative approach, it is tempting to try to use the EM density given by the EM methodology to construct an approximate 95% confidence ellipsoid for $\mathbf{M}(p)$ directly. The problem with this approach is that the large variability of the higher moments can lead to estimates that are much less likely to be below the EM estimate. This sort of approach would only be appropriate for sufficiently large sample sizes.



	mean	var	skew	flat
p	0	1	0	2
q	0	1	0	3

Fig. 2. *Two moments*: The first series of boxplots show the distribution of 500 sample EM estimates over a range of increasing sample sizes. The second and third series of boxplots show the corresponding AM and CL sample utilities described in Sections 4 and 5. For each sample size, 500 independent ensembles were generated. Each boxplot shows the range of the results for the 500 ensembles with the box representing the middle 50%, and the horizontal line through the box representing the median value. The horizontal dashed line across all the boxplots represents the EM estimate.

The adjusted moment methodology tends to satisfy the desired properties in Definition 1. However, it is not very adept at detecting information in the higher moments. This methodology seems to perform best for the two-moment case. The adjusted four-moment estimates in Figs. 3 and 6 do not show significant improvements over the two-moment counterparts in Figs. 2 and 5 for the smaller sample sizes. In addition, the adjusted non-Gaussianity estimates in Figs. 4 and 7 are zero or close to zero, even for larger sample



	mean	var	skew	flat
p	0	1	0	2
q	0	1	0	3

Number of ensembles: 500

Relative entropy: 0.06016

Fig. 3. *Four moments*: The first series of boxplots show the distribution of 500 sample EM estimates over a range of increasing sample sizes. The second and third series of boxplots show the corresponding AM and CL sample utilities described in Sections 4 and 5. For each sample size, 500 independent ensembles were generated. Each boxplot shows the range of the results for the 500 ensembles with the box representing the middle 50%, and the horizontal line through the box representing the median value. The horizontal dashed line across all the boxplots represents the EM estimate.

sizes. On the other hand, the series of boxplots in Figs. 2 and 5 demonstrate the effectiveness of this methodology for the two-moment case. For the first figure, where the EM estimate is zero, the adjusted moment methodology yields very few nonzero estimates. This is an important special case, since the EM methodology is the most biased for this case. In the second figure, the EM estimate lies fairly close to the 95% mark for the various data sets, as desired.

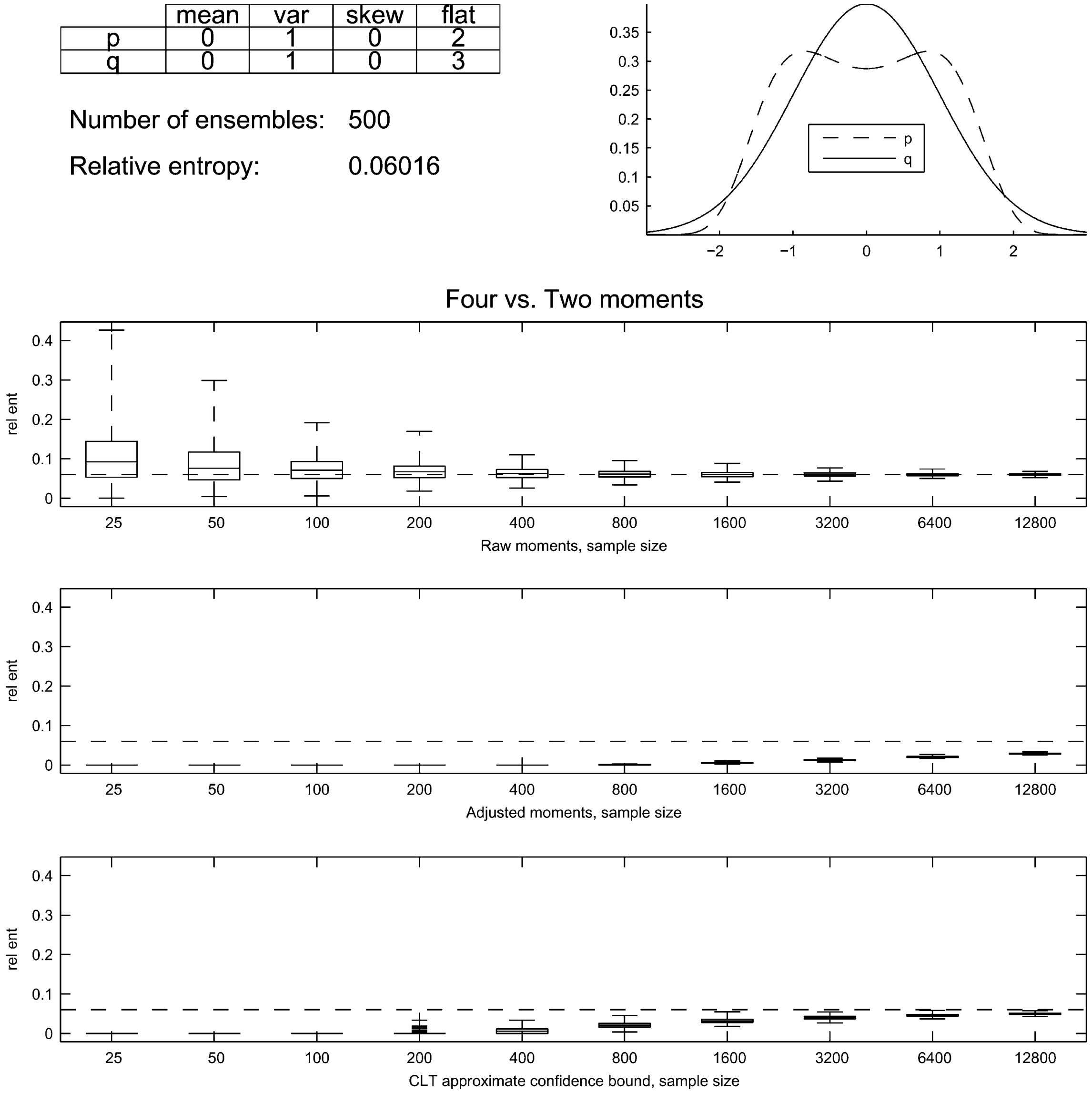


Fig. 4. *Non-Gaussianity*: The first series of boxplots show the distribution of 500 sample EM estimates over a range of increasing sample sizes. The second and third series of boxplots show the corresponding AM and CL sample utilities described in Sections 4 and 5. For each sample size, 500 independent ensembles were generated. Each boxplot shows the range of the results for the 500 ensembles with the box representing the middle 50%, and the horizontal line through the box representing the median value. The horizontal dashed line across all the boxplots represents the EM estimate.

A more rigorous explanation of the adjusted moment procedure as well as the statement and proof of the necessary central limit theorem required by this theory are given in Appendix A. In the following section, a different measure of sample utility is proposed that uses a central limit result for P that is proved in Appendix B.

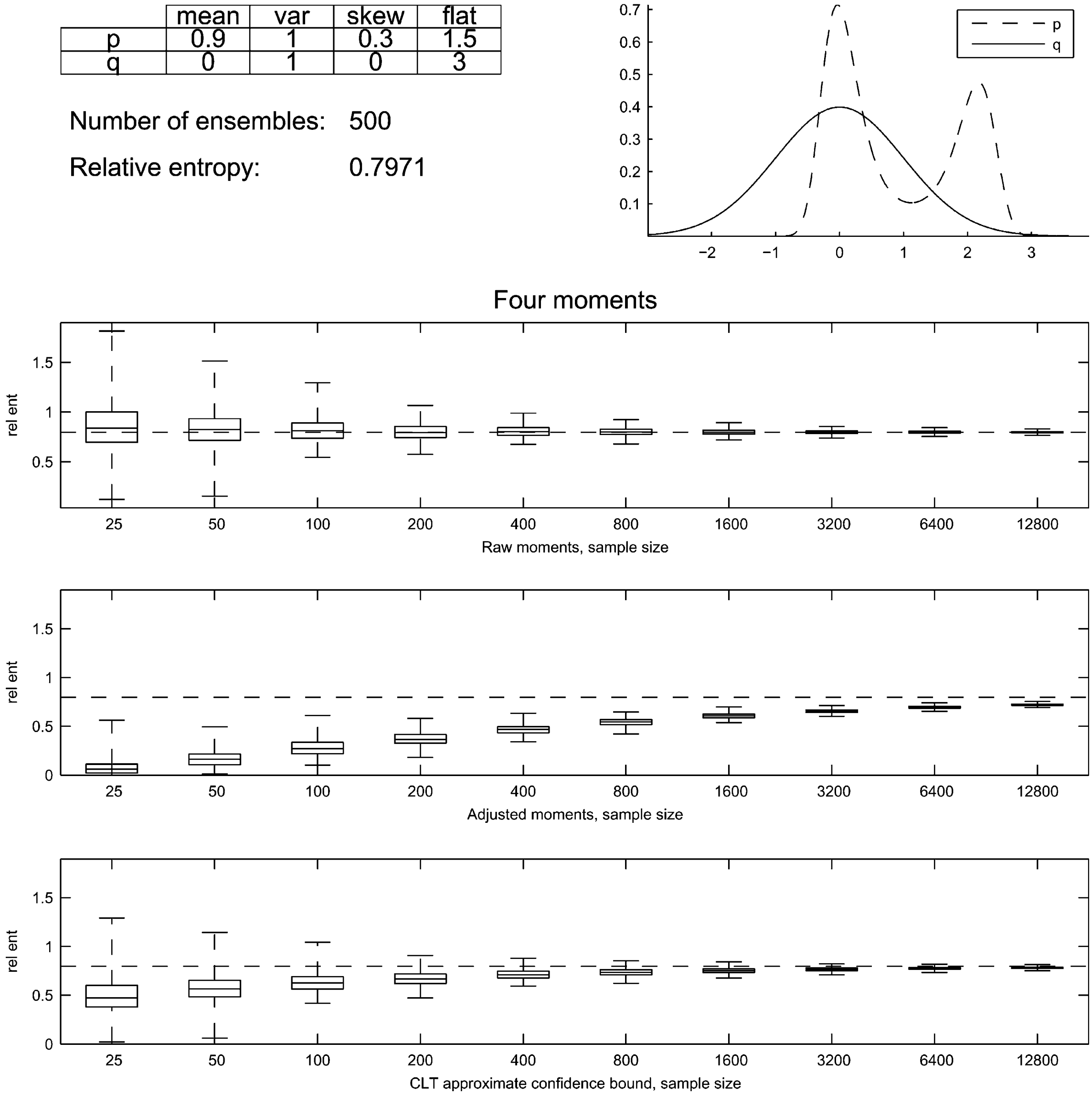


Fig. 5. *Two moments*: The first series of boxplots show the distribution of 500 sample EM estimates over a range of increasing sample sizes. The second and third series of boxplots show the corresponding AM and CL sample utilities described in Sections 4 and 5. For each sample size, 500 independent ensembles were generated. Each boxplot shows the range of the results for the 500 ensembles with the box representing the middle 50%, and the horizontal line through the box representing the median value. The horizontal dashed line across all the boxplots represents the EM estimate.

5. Central limit theorem methodology

In this section, a measure of sample utility is defined directly in terms of a 95% confidence interval for the sample EM estimate itself. The procedure depends upon a central limit result for P, which is stated and proved in Appendix B.

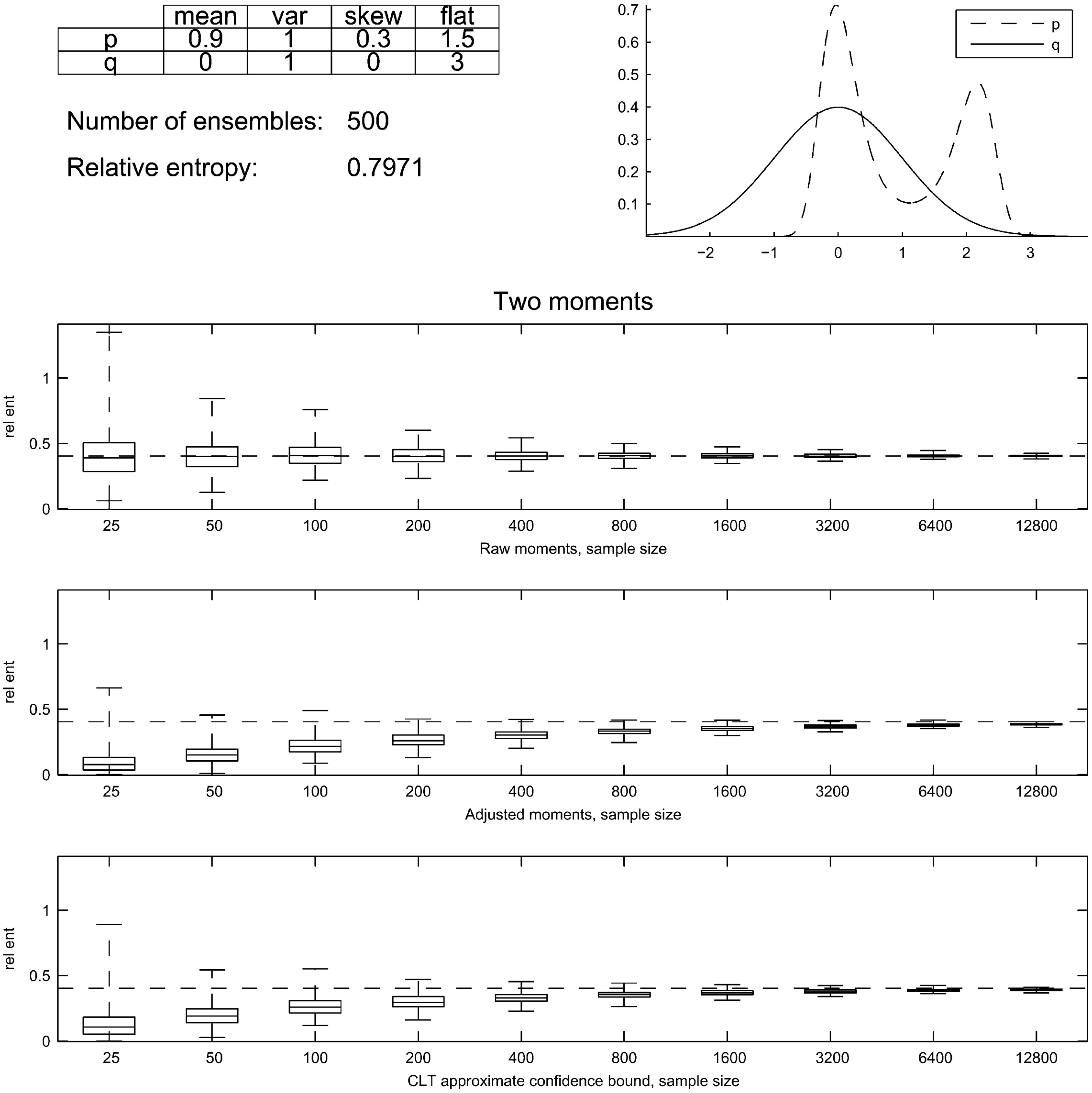
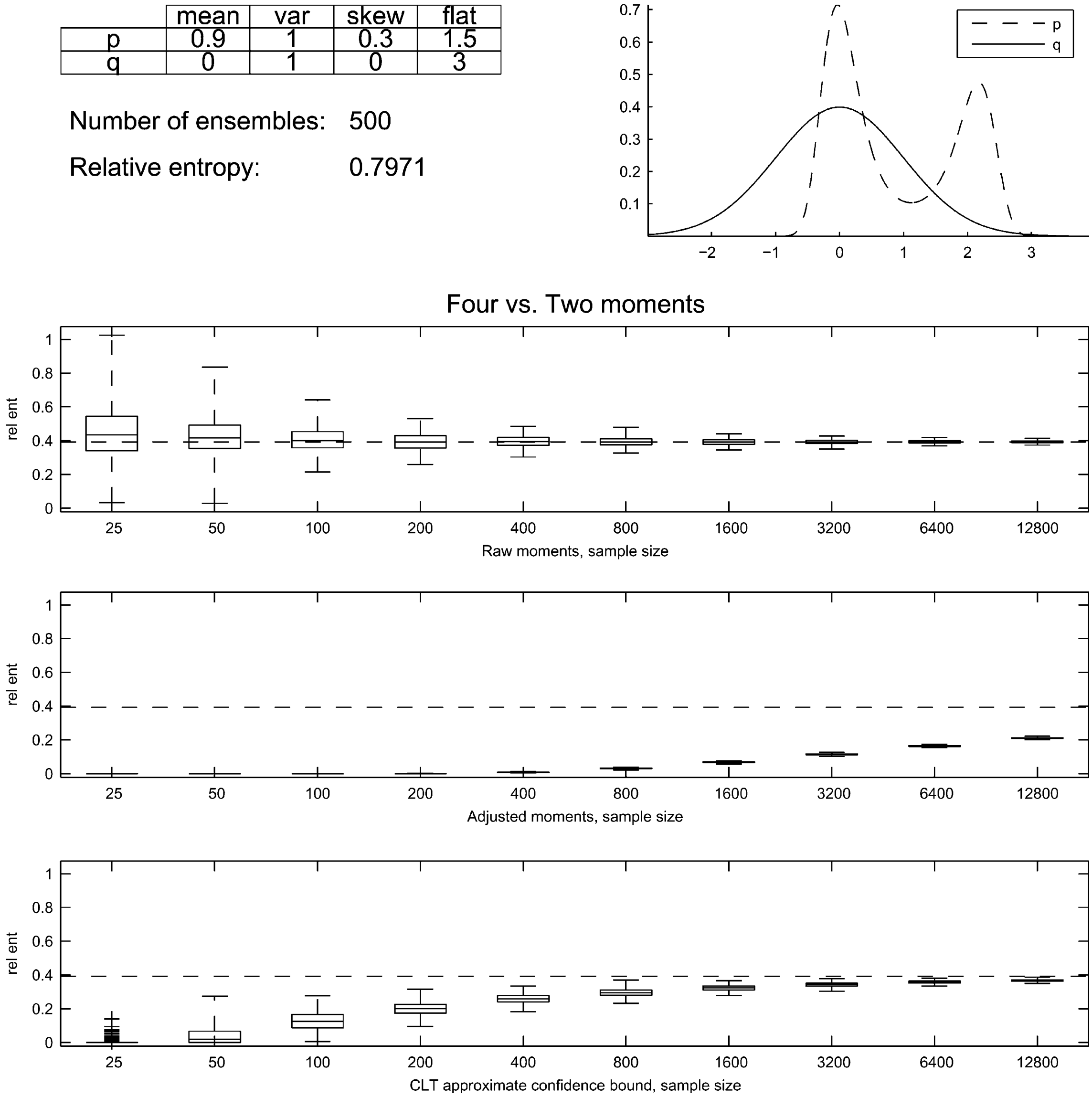


Fig. 6. *Four moments*: The first series of boxplots show the distribution of 500 sample EM estimates over a range of increasing sample sizes. The second and third series of boxplots show the corresponding AM and CL sample utilities described in Sections 4 and 5. For each sample size, 500 independent ensembles were generated. Each boxplot shows the range of the results for the 500 ensembles with the box representing the middle 50%, and the horizontal line through the box representing the median value. The horizontal dashed line across all the boxplots represents the EM estimate.

The central limit result states that, as sample size approaches infinity, $P(\mathbf{s}|\mathbf{M}(q))$ will approach a Gaussian distribution centered about $P(\mathbf{M}(p)|\mathbf{M}(q))$. In Theorem B.4 in Appendix B, we give an explicit formula for the variance of this Gaussian distribution which involves a complicated explicit formula depending on the moment hierarchy. This theoretical fact leads to the following.



	mean	var	skew	flat
p	0.9	1	0.3	1.5
q	0	1	0	3

Fig. 7. *Non-Gaussianity*: The first series of boxplots show the distribution of 500 sample EM estimates over a range of increasing sample sizes. The second and third series of boxplots show the corresponding AM and CL sample utilities described in Sections 4 and 5. For each sample size, 500 independent ensembles were generated. Each boxplot shows the range of the results for the 500 ensembles with the box representing the middle 50%, and the horizontal line through the box representing the median value. The horizontal dashed line across all the boxplots represents the EM estimate.

5.1. EM central limit algorithm

As is commonly done in the statistical literature [28], all the parameters in the formula for the variance that are unknown are approximated by the corresponding finite sample quantities. Once the approximate variance is known, an approximate, one-sided, 95% confidence interval for the EM estimate can be constructed. The lower bound cutoff for the confidence interval is $P(\mathbf{s}|\mathbf{M}(q)) - 1.965 * \hat{\sigma}/N$, where $\hat{\sigma}^2$ is the approximate variance and N is the sample size. The value 1.965 is the value for which a 95% of the weight of standard Gaussian density lies below. This lower bound cutoff is defined to be the central limit (CL) sample utility. If the confidence cutoff is less than zero, the CL sample utility is defined to be zero. Intuitively, this algorithm has the desirable features in Definition 1 at least for sufficiently large sample sizes. For sufficiently large sample sizes, the EM estimate lies above the CL sample utility, with a 95% level of confidence. As the sample size increases, the CL sample utility will increase to the EM estimate. For small sample sizes, the 95% confidence level still can remain fairly accurate since in statistics approximate confidence intervals are often quite accurate for small sample sizes [28].

In the final panel of Figs. 2–7, a similar series of boxplots to the third and fourth panels are shown for this central limit theorem methodology. Since the sample size is finite and the variance is approximated, it is surprising that, even for relatively small sample sizes, the lower bound cutoff from the central limit theorem seems to satisfy the three properties listed in Definition 1. In addition, this methodology is much better than the adjusted moment methodology at detecting information content due to the higher moments. Comparing Fig. 2 to Fig. 3, even at sample size 25, the four-moment estimates are able to successfully detect the larger information content due to the higher moments. This is especially surprising, given the closeness of p to the standard Gaussian q. The ability of this methodology to detect higher moment information is even more pronounced in Figs. 5 and 6. The detection of non-Gaussianity, in Fig. 4, is also better than in the adjusted moment methodology. In the first figure, as expected, there is little detection of higher moment information until larger sample sizes. In the second figure, non-Gaussianity information begins to be detected at sample size 50. Thus, non-Gaussianity information can be detected with confidence at small sample sizes.

In the following section, the methodologies introduced in Sections 4 and 5 are applied to the Lorenz ‘96 model in an attempt to detect non-Gaussian tendencies in ensemble prediction densities, with 95% confidence.

6. Application to the Lorenz ‘96 model

In this section, the adjusted moment and central limit algorithms for detection of non-Gaussianity, introduced in Sections 4 and 5, are carried out for the damped forced Lorenz ‘96 model.

The Lorenz ‘96 model is the spatially discrete family of equations given by

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} = (u_{j+1} - u_{j-2})u_{j-1} - u_j + F, \quad j = 0, 1, \ldots, J-1 \tag{4}$$

with periodic boundary conditions, $u_0 = u_J$. The term $-u_j$ in (4) represents damping (with a unit time scale of 5 days) while F represents constant “solar forcing” (see [25,15]). The model in (4) is designed to mimic midlatitude weather and climate behavior, so periodic boundary conditions are appropriate. The unit spatial scale between discrete nodes is regarded as a nondimensional midlatitude Rossby radius ≈800 km and for this reason the discrete system size is set to be $J = 40$ nodes. In midlatitude weather systems, the main “weather waves”, the Rossby waves have westward (toward negative x) phase velocity, but from out own anecdotal experience, weather systems collectively move eastward (toward positive x) with unstable behavior. The models in (4) have analogous behavior. The modes of the system are discrete Fourier modes with wavenumbers k ranging $-20 < k \leqslant 20$. Analogous to real weather systems, the models produce bands of unstable waves centered about the wavenumber $|k| = 8$ with westward phase velocities and overall eastward group velocities, and have strongly chaotic dynamics. These results, as well as a more detailed description of the model, can be found in [12]. In the current paper, the robust dynamical regime is used with constant forcing $F = 8$ and damping coefficient $d = 1$. Below we examine the non-Gaussian information content in ensemble predictions with fairly small sample sizes for the three Fourier modes, $k = 0$, $k = 3$ and $k = 8$; the mode $k = 0$ defines the climatology, while $k = 8$ is the most unstable mode in the model and $k = 3$ is a linearly stable but chaotic large scale mode (see [12]).

The numerical set-up for the experiments with the Lorenz '96 model is the following: first, a long (10,000 time units) "climatological" time series of a single solution is generated. Then an instantaneous snapshot of this solution is recorded at the end of the series. The statistical ensembles of various sizes are then bred around this single recorded snapshot, by perturbing each gridpoint via a narrow Gaussian probability with small variance (10^{-5} fraction of the climatological variance). The ensembles then propagate further, as their time series are being recorded.

Four different ensemble sizes with 25, 50, 100 and 200 members are employed in the current work to evaluate and compare the efficiency of the methodologies for measuring information content for different sample sizes. Since an ensemble provides a single value of the relative entropy at any given time, a super-ensemble (ensemble of ensembles) has to be generated for each of the four sample sizes in order to study

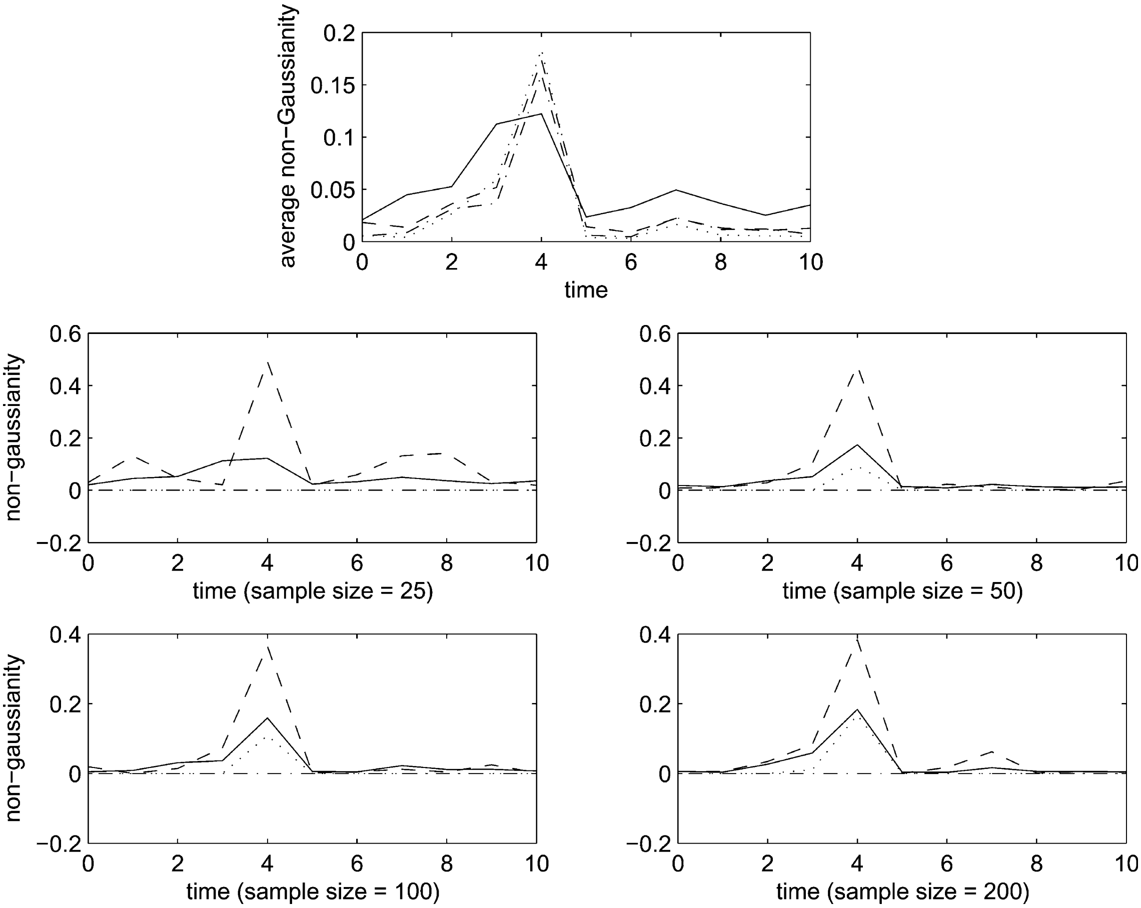


Fig. 8. Fourier mode: $\mathbf{k} = \mathbf{0}$. Top panel shows the average non-Gaussianity, over 20 ensembles, with sample sizes 25 (solid), 50 (dashed), 100 (dot-dashed), and 200 (dotted). Panels 2–4, the solid curve shows the average non-Gaussianity for the indicated sample size. The dashed curve shows the ensemble with the largest non-Gaussianity at time $t = 4$. The dot-dashed curve shows the adjusted moment technique applied to this ensemble. The dotted curve shows the central limit technique applied to this ensemble.

meaningful statistical properties of non-Gaussianity in different regimes. For each sample size, a super-ensemble consisting of 20 ensembles is employed.

The average of the sample non-Gaussianities of the Fourier mode $k = 0$ are plotted in the first panel of Fig. 8, with sample sizes 25 (solid), 50 (dashed), 100 (dot-dashed), and 200 (dotted). In general, the average of the sample non-Gaussianities is not equal to the true non-Gaussianity, but here it is assumed. The graphs show the averages decreasing with increasing sample size except for a spike in the non-Gaussianity at time $t = 4$. This behavior is counter to the desired behavior described in Definition 1.

The second panel of Fig. 8 shows the average sample non-Gaussianity for sample size 25 as a solid curve. The dashed curve represents the ensemble with the largest non-Gaussianity at time $t = 4$. The dot-dashed and dotted curves, which overlap on this graph, show the adjusted moment and central limit methodologies, respectively, applied to this ensemble. Neither method produces any non-Gaussianity information, due to the small sample size. Similar graphs are shown for each sample size in panels three, four, and five. In each panel, the solid curve represents the average sample non-Gaussianities, while the dashed, dot-dashed, and dotted curves represent the sample EM, adjusted moment, and central limit methodologies applied to the ensemble with the largest non-Gaussianity at time $t = 4$. As expected from the earlier results in Section

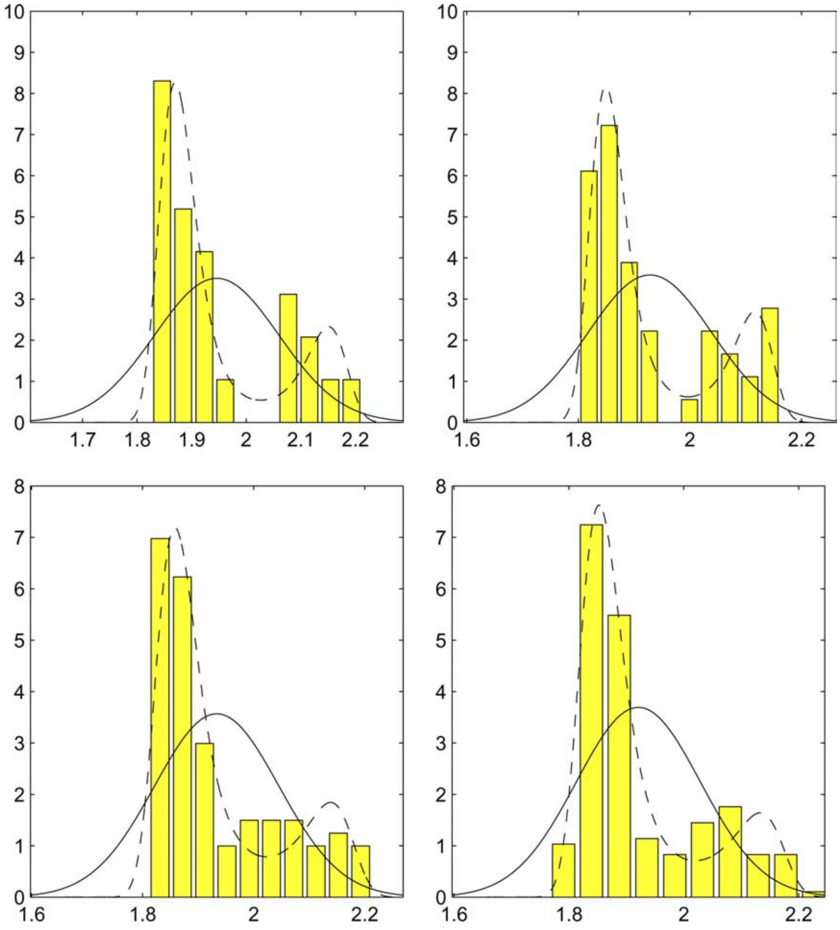


Fig. 9. Fourier mode: $\mathbf{k} = \mathbf{0}$. Each graph shows a histogram of the data from panels 2–4 in Fig. 8 at time $t = 4$. The Gaussian and four-moment fits are overlayed.

4, the adjusted moment methodology is not effective at detecting the non-Gaussianity. On the other hand, the central limit theorem method produces a curve below and increasing toward the average sample non-Gaussianity curve. The plots demonstrate that evidence of non-Gaussian behavior of the prediction density can be detected, with 95% confidence, at sample sizes as small as 50. Fig. 9 shows the histograms of the data, at each sample size and time $t = 4$, for the ensemble with the maximum non-Gaussianity at time $t = 4$. The solid curve shows the Gaussian sample EM density, and the dashed curve shows the four-moment sample EM density. Clearly, a four-moment density fits the data better than the Gaussian density. However, the issue here is whether the Gaussian density could have produced the data.

Figs. 10–13 show the same plots as Figs. 8 and 9 for Fourier modes $k = 3$ and $k = 8$. In each case, the observed ensemble is the one with maximum non-Gaussianity in Fourier mode $k = 0$. For these figures, the histograms are plotted for the time $t = 3$, where the largest peak of non-Gaussianity seems to appear. Here again, the central limit methodology detects non-Gaussian behavior of the prediction density, with 95% confidence.

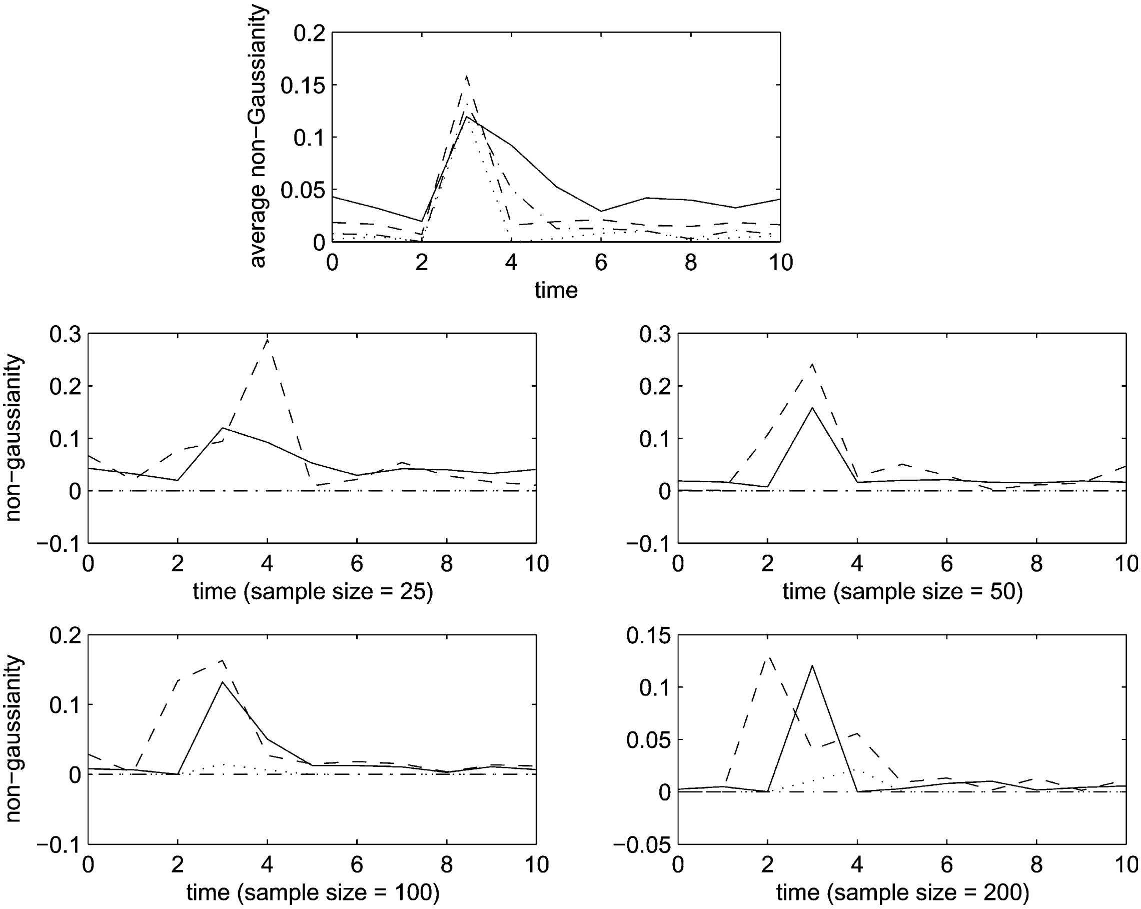


Fig. 10. Fourier mode: **k = 3**. Top panel shows the average non-Gaussianity, over 20 ensembles, with sample sizes 25 (solid), 50 (dashed), 100 (dot-dashed), and 200 (dotted). Panels 2–4, the solid curve shows the average non-Gaussianity for the indicated sample size. The dashed curve shows the ensemble with the largest non-Gaussianity in the $k = 0$ mode at time $t = 3$. The dot-dashed curve shows the adjusted moment technique applied to this ensemble. The dotted curve shows the central limit technique applied to this ensemble.

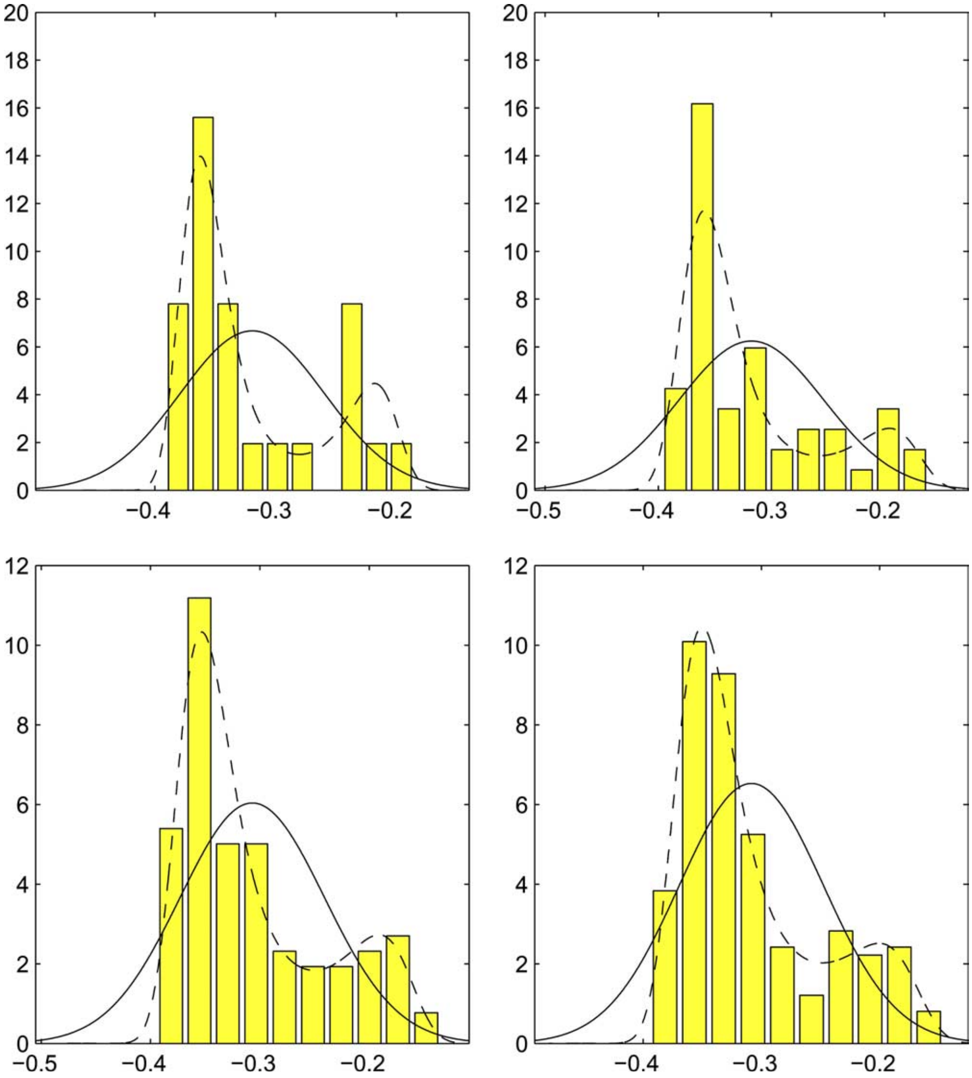


Fig. 11. Fourier mode: **k = 3**. Each graph shows a histogram of the data from panels 2–4 in Fig. 10 at time $t = 3$. The Gaussian and four-moment fits are overlayed.

An interesting example of a false indication of bimodality for the most unstable mode, $k = 8$ can be seen in Fig. 13. In the first panel, at sample size 25, there seems to be a strong indication of non-Gaussianity. Visually, the data seems bimodal, and the corresponding sample EM estimate, the dashed line in panel two of Fig. 12 at $t = 3$, is ≈ 0.5. However, at a sample size of 200, in the fourth panel, the bimodality is almost completely gone, both visually and by sample EM estimate. Interestingly, the central limit methodology detects little or no non-Gaussian behavior for all sample sizes, and this is a highly desirable feature.

The results suggest, with 95% confidence, that non-Gaussian tendencies exist for some modes of the prediction densities at various times. That the higher moment information is detected, with confidence and for small sample sizes, is evidence that a strictly Gaussian prediction strategy will sometimes miss substantial levels of predictability.

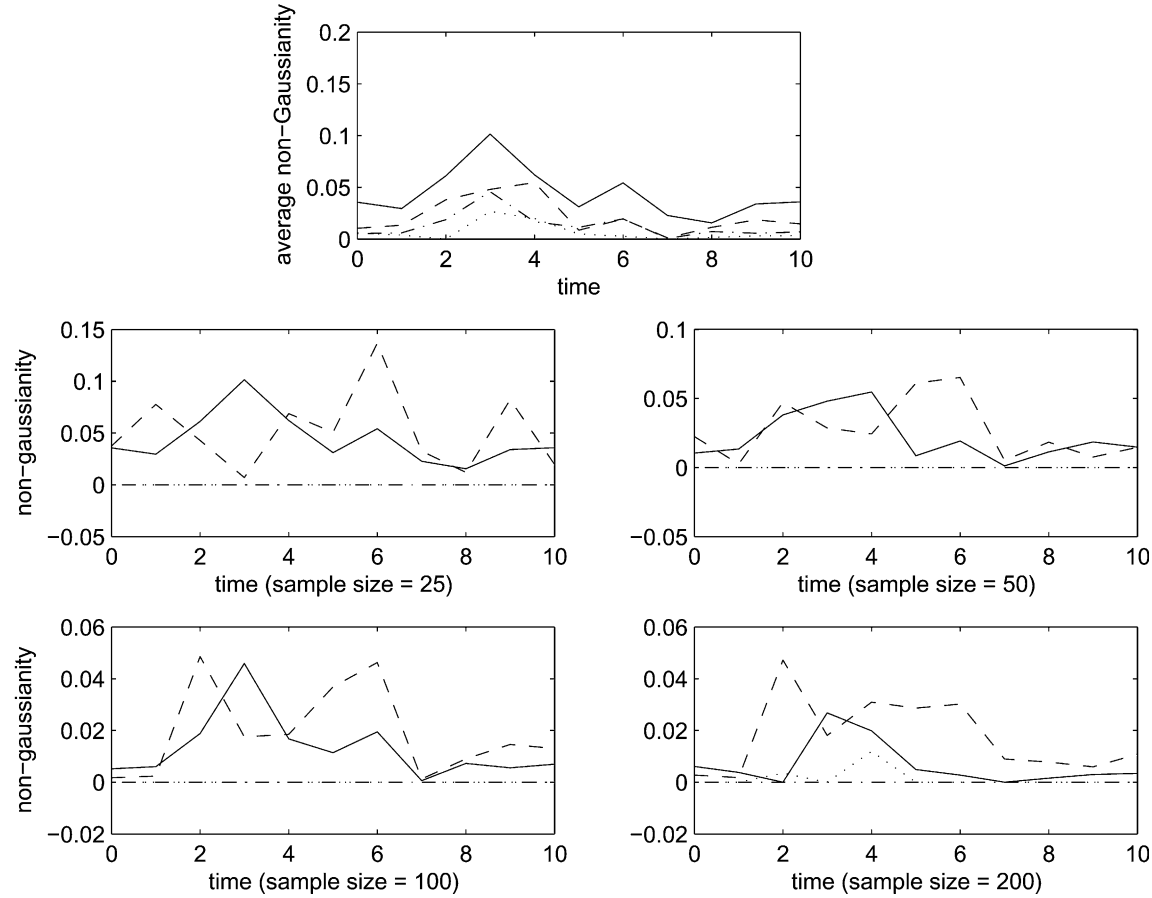


Fig. 12. Fourier mode: **k** = **8**. Top panel shows the average non-Gaussianity, over 20 ensembles, with sample sizes 25 (solid), 50 (dashed), 100 (dot-dashed), and 200 (dotted). Panels 2–4, the solid curve shows the average non-Gaussianity for the indicated sample size. The dashed curve shows the ensemble with the largest non-Gaussianity in the $k = 0$ mode at time $t = 3$. The dot-dashed curve shows the adjusted moment technique applied to this ensemble. The dotted curve shows the central limit technique applied to this ensemble.

7. Summary

The information theory framework utilizing relative entropy is extended here to estimate uncertainties in predictions coming from the limited sample size of a forecast ensemble. Two methodologies are devised to compensate for the lack of information due to small sample size: one is based on a null-hypothesis testing for general non-Gaussian moments of the probability density functions and leads to the Adjusted Moment Algorithm of Section 4, while the other employs a central limit theorem for the moment-based relative entropy itself and leads to the EM central limit algorithm of Section 5. The two methodologies are systematically tested against the straightforward "perfect predictability" EM method through two series of experiments involving both the explicitly defined family of non-Gaussian probability density functions with parameterized skewness and bimodality, and the Lorenz '96 model, a simple truncation of the Burgers–Hopf equation with damping and constant forcing. The Lorenz '96 model was picked for its simplicity

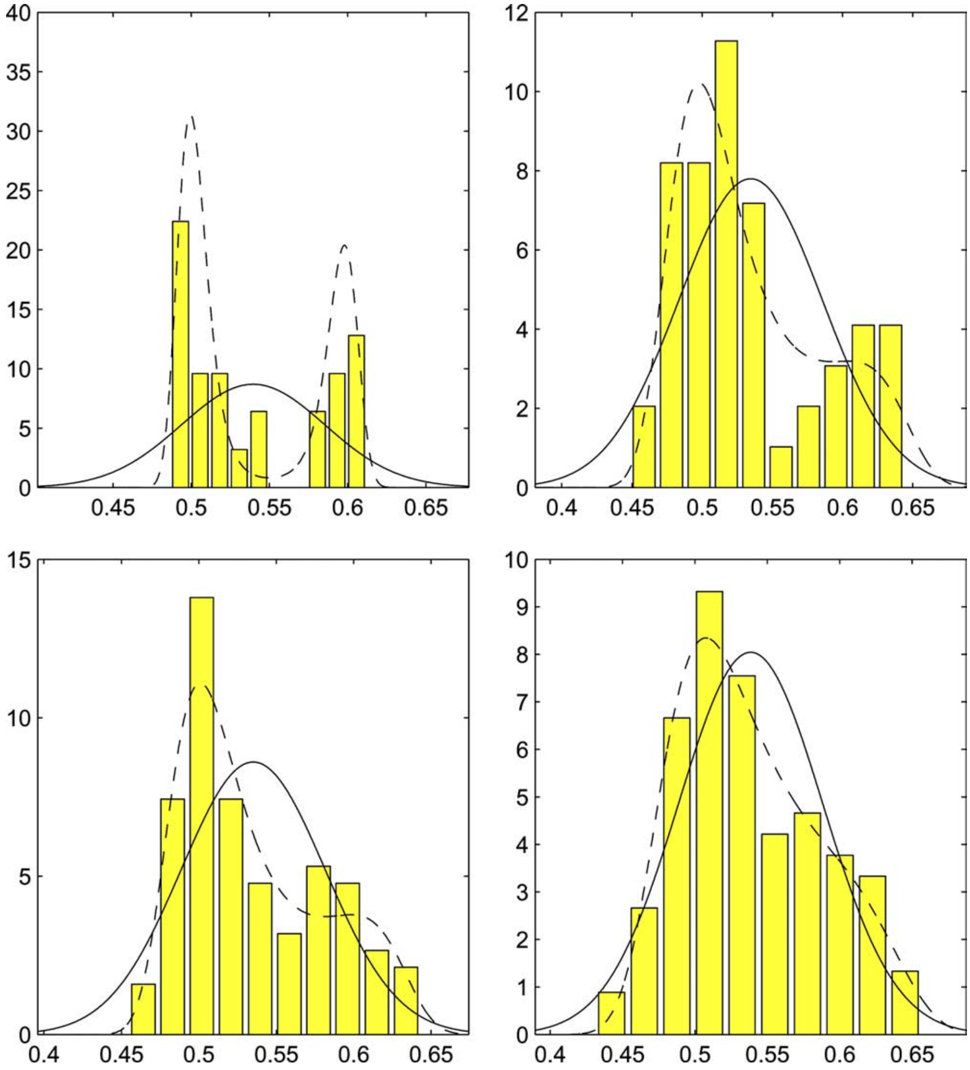


Fig. 13. Fourier mode: **k = 8**. Each graph shows a histogram of the data from panels 2–4 in Fig. 12 at time $t = 3$. The Gaussian and four-moment fits are overlayed.

and the ability to mimic the chaotic behavior of unstable wave patterns in complex weather systems. We summarize the results of the study in the following remarks:

- *Perfect predictability methodology.* Even though the perfect predictability methodology is not a novel technique developed here and has been used in previous ensemble prediction studies only with large sample sizes, the authors summarize the trends here for easy comparison with other methodologies. The perfect predictability methodology has the practical advantage that it is theoretically centered at the EM estimate for information. Two qualitatively different types of behavior can be distinguished for this method with varying sample size: first, when the actual relative entropy of two probability density functions is large, the perfect predictability methodology overshoots and undershoots the truth with roughly equal possibility for all considered sample sizes; and second, when the actual relative entropy is small, this straightforward method tends to overshoot the truth, gradually adjusting to it with increased sample size. Both types of behavior are expected: the latter is due to the fact that in the limiting case of $p = q$ (zero utility) an undersampled p can only produce overshoots (relative entropy can not be negative), while the former occurs because for large relative entropy both undershoots and overshoots are roughly equally possible.

- *Adjusted moments methodology.* For the two-moment (Gaussian) estimates the adjusted moments methodology shows significant improvement over the perfect predictability methodology for small values of relative entropy. There is no improvement for large values of relative entropy, because in that case the results of perfect predictability methodology are statistically centered at the truth, while the adjusted moments methodology systematically undershoots the truth (this is what it has been designed for in the first place). But even for systematic undershoots, the adjusted moment methodology shows rapid convergence to the truth with increasing sample size. There is no significant improvement, however, for the four-moment estimates over two-moment estimates, which seems to be an intrinsic deficiency of the adjusted moments strategy.
- *Central limit theorem methodology.* While the general behavior of the central limit technique has trends similar to those of the adjusted moments methodology, there is significant improvement compared to adjusted moments methodology. First, the information from the non-Gaussian moments is now detected successfully, which is also practically demonstrated in the Lorenz '96 model with sample sizes in the range 50–100. Second, the convergence to the true relative entropy with increasing sample size is generally much faster than that for the adjusted moments methodology. Third, the method is robust at small sample sizes to false detection of non-Gaussianity. It is therefore concluded that the central limit theorem methodology is superior to the adjusted moments methodology, at least for purposes of the current paper, and within its testing framework.
- *Generalization for multivariate distributions.* All three algorithms discussed above generalize in a straightforward fashion conceptually to multivariate distributions in N variables. However, there are major practical computational issues in this setting. First, there are no practical ways to compute the relative entropy in (1) directly through quadrature for $N \gg 1$ even if the distributions p,q are known exactly. Secondly, even for the perfect predictability algorithm, the optimization procedure for four-moment constraints becomes prohibitively expensive for large N. The approach utilized by Abramov and Majda [12] in this setting to use mathematical theory [10] to bound from below the information content in (1) for large N by a sum of N one-dimensional relative entropies plus a sum of $N(N + 1)/2$ two-dimensional entropies. A rapid optimization algorithm for four moment estimators for one- and two-dimensional relative entropies is developed and used in [12,16] in conjunction with this approach (see [29] for algorithmic details). This algorithmic strategy can be applied directly to both the adjusted moment and central limit statistical methodologies. However, the interesting issue of the sample sizes for significant non-Gaussian detection through the algorithms for the two-dimensional distributions remains to be explored.

Overall, the current work demonstrates the general versatility of the finite sample estimates. It is shown that the methodologies are devised under the rigorous framework of the information theory and based upon the fact that finite sample sizes reduce the information content in terms of relative entropy. The "lab" testing with explicitly given probability densities reveals the consistency of the methodologies and their estimates with the exact relative entropy measurements. Finally, the "field" testing, based upon the Lorenz '96 model, unambiguously shows the practical applicability of the small sample estimation strategies developed here. Clearly, the EM central limit algorithm is a promising one for detecting non-Gaussianity in practical ensemble predictions with relatively small sample size.

Finally, in [14], the term sample utility is used to refer to a particular sample estimate of the EM estimate which uses a Bayesian approach. In this approach, the unknown PDF, p, is assumed to belong to a certain family of distributions with undetermined parameters. The undetermined parameters are assumed to be random, with a uniform prior distribution. Conditioning on the sample moments leads to a posterior distribution, which is used to compute the expected information loss from using a family member with moments equal to the sample moments. This expected information loss is assumed to be the information loss due to sample estimation, and thus, subtracted from the sample EM estimate. This approach is a formal attempt to eliminate the bias in using the sample EM estimate and not an attempt to define a quantity possessing the properties suggested in Definition 1. Since the goal of this approach is very different from the goal of the approaches taken in this paper, these strategies should not be compared.

Acknowledgments

The authors thank Richard Kleeman for his advice and help on many of the issues discussed in this paper. Kyle Haven is partially supported as an NSF-VIGRE postdoctoral fellow at the Courant Institute. Andrew Majda is partially supported by the NSF Grant DMS-9972865, the ONR Grant N00014-96-1-0043, and the NSF CMG Grant DMS-0222133. Rafail Abramov is supported as a postdoctoral fellow through these research grants.

Appendix A. Details of adjusted moment methodology

Here, the mathematical details that support the discussion in Sections 3 and 4 are given. The exact definitions of the involved quantities are given, followed by the statements and proofs or the needed results.

For any probability density, ρ, the mean and centered moments of ρ are defined as

$$M_1(\rho) = \int x\rho(x)\,\mathrm{d}x \quad \text{and} \quad M_k(\rho) = \int (x - M_1(\rho))^k \rho(x)\,\mathrm{d}x, \quad k \geqslant 2. \tag{A.1}$$

Throughout the discussion, it is assumed that q is of the form

$$q(x) = \exp\left[-\sum_{i=0}^{K} \alpha_i (x-\mu)^i\right], \tag{A.2}$$

where $\mu = M_1(q)$ is the mean of q and $\boldsymbol{\alpha}$ are uniquely determined by the moments $\mathbf{M}(q)$. For this paper, the value of K is either 2 or 4.

When $\mathbf{M}(p)$ is known precisely, a lower bound estimate of $R(p|q)$ is found by minimizing $R(\rho|q)$ over all probability densities satisfying $\mathbf{M}(\rho) = \mathbf{M}(p)$. The convexity of R ensures that the minimum will be reached for some p_K, the entropy moment (EM) PDF. The EM PDF has the form

$$p_K(x) = Z(\boldsymbol{\theta}, v)^{-1} \exp\left[\theta_1 x + \sum_{k=2}^{K} \theta_k (x-v)^k\right] q(x), \tag{A.3}$$

where $\boldsymbol{\theta}$ are the Lagrange multipliers for the moment constraints, $v = M_1(p)$, and

$$Z(\boldsymbol{\theta}, v) = \int \exp\left[\theta_1 x + \sum_{k=2}^{K} \theta_k (x-v)^k\right] q(x)\,\mathrm{d}x$$

is the normalizing constant. When $K = 2$ and q is Gaussian, the EM density is Gaussian.

The EM lower bound estimate of $R(p|q)$ is defined as $R(p_K|q)$, which is given by

$$R(p_K|q) = -\log(Z(\boldsymbol{\theta}, v)) + \boldsymbol{\theta} \cdot \mathbf{M}(p). \tag{A.4}$$

It follows immediately from the definition of p_K that $R(p|q) \geqslant R(p_K|q)$. The definition of the minimum information content given only $\mathbf{M}(p)$ is given by

$$P(\mathbf{M}(p)|\mathbf{M}(q)) = \min\{R(\rho|q) : \mathbf{M}(\rho) = \mathbf{M}(p)\} = -\log(Z(\boldsymbol{\theta}, v)) + \boldsymbol{\theta} \cdot \mathbf{M}(p). \tag{A.5}$$

Let N denote the sample size and $x_1,\ldots,x_N$ denote the observed sample data from p. The sample data can be thought of as just one realization from the set $X_1,\ldots,X_N$ of iid p-distributed random variables. The sample moments of the random variables are given by

$$S_1 = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{and} \quad S_k = \frac{1}{N}\sum_{i=1}^{N}(X_i - S_1)^k, \quad k \geqslant 2. \tag{A.6}$$

The sample moments, $\mathbf{s} = (s_1,\ldots,s_K)$, are defined as in Eq. (A.6) with $\mathbf{x}$ replacing $\mathbf{X}$ and can be thought of as just one realization of the random vector $\mathbf{S} = (S_1,\ldots,S_K)$.

We are now ready to state and prove the central limit result needed for the definition of the approximate confidence ellipsoids, described in Section 4, and the approximate variance of $P(\mathbf{S}|\mathbf{M}(q))$, described in Section 5. This proposition is also needed for the proof of Theorem B.4. Both Proposition A.3 and Theorem B.4 rely heavily upon a standard result known as Slutsky's Theorem, which for convenience, is reprinted here.

Theorem A.2 (Slutsky). *If $g(x,y)$ is a jointly continuous function, X_n converges to X in distribution, and Y_n converges to a constant a in probability; then $g(X_n, Y_n)$ converges to $g(X,a)$ in distribution.*

Proposition A.3. *Assume that $X_1,\ldots,X_N$ are iid random variables from a common distribution with density q. Then $\sqrt{N}(\mathbf{S} - \mathbf{M}(q))$ converges to a Gaussian distribution with mean zero and covariance $C = A\Sigma A^{\mathrm{T}}$, where*

$$\Sigma_{i,j} = \int ((x-\mu)^i - M_i(q))((x-\mu)^j - M_j(q))q(x)\,\mathrm{d}x, \tag{A.7}$$

where A is a 2×2 identity matrix for $K = 2$, and

$$A = I_{4\times 4} - \sum_{k=3}^{K} kM_{k-1}(q)\mathbf{e}_k \tag{A.8}$$

for $K = 4$. Here, $I_{4\times4}$ denotes a 4×4 identity matrix, and $\mathbf{e}_k$ denotes the kth standard unit vector. In the definition of Σ, in Eq. (A.7), $M_1(q)$ is defined to be zero.

Proof. The proof is carried out for the case when $K = 4$. The proof for the $K = 2$ case is similar and simpler. First, consider the asymptotics of the sample moments centered around μ

$$\hat{M}_1 = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{and} \quad \hat{M}_k = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^k, \quad k \geqslant 2. \tag{A.9}$$

The central limit theorem states that $\sqrt{N}(\hat{\mathbf{M}}(q) - \mathbf{M}(q))$ converges to a Gaussian distribution with mean zero and covariance given by Σ in Eq. (A.7).

In order to determine the asymptotics of $\sqrt{N}(\mathbf{S} - \mathbf{M}(q))$, Slutsky's theorem is needed. For $k \geqslant 2$

$$S_k = \frac{1}{N}\sum_{i=1}^{N}(X_i - S_1)^k = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{k}\binom{k}{j}(X_i-\mu)^j(\mu - S_1)^{k-j} = \sum_{j=2}^{k}\binom{k}{j}\hat{M}_j(\mu - S_1)^{k-j} + (1-k)(\mu - S_1)^k.$$

Thus

$$\sqrt{N}(S_k - M_k(q)) = \sqrt{N}(\hat{M}_k - M_k(q)) + \sum_{j=2}^{k-1}\binom{k}{j}\hat{M}_j(\sqrt{N}(\mu - S_1))^{k-j}N^{\frac{1-k+j}{2}} + (1-k)(\sqrt{N}(\mu - S_1))^k N^{\frac{1-k}{2}}. \tag{A.10}$$

Since $\sqrt{N}(\mu - S_1)$ converges to a Gaussian random variable and $N^{\frac{1-k}{2}}$ converges to zero, Slutsky's theorem implies that the last term in Eq. (A.10) converges to zero. Similarly for $j < k-1$, since $\hat{\mathbf{M}}$ converge to $\mathbf{M}(q)$, and $N^{\frac{1-k+j}{2}}$ converges to zero, all but the $(k-1)$th term in the sum will converge to zero. Therefore

$$\sqrt{N}(\mathbf{S} - \mathbf{M}(q)) = \sqrt{N}(\hat{\mathbf{M}} - \mathbf{M}(q)) + \sum_{k=3}^{4} k\hat{M}_{k-1}\sqrt{N}(\mu - S_1) + R_N = \hat{A}[\sqrt{N}(\hat{\mathbf{M}} - \mathbf{M}(q))] + R_N,$$

where $\hat{A}$ is the approximation of A in Eq. (A.8) and R_N is a remainder term that converges to zero in distribution. Since $\sqrt{N}(\hat{\mathbf{M}} - \mathbf{M}(q))$ converges to a Gaussian distribution with covariance Σ and $\hat{A}$ converges to A, it follows from Slutsky's Theorem and the last display that $\sqrt{N}(\mathbf{S} - \mathbf{M}(q))$ converges to a Gaussian distribution with covariance C. This ends the proof of the proposition.

Although $P(\mathbf{s}|\mathbf{M}(q))$ approaches $P(\mathbf{M}(p)|\mathbf{M}(q))$ as $N \to \infty$, it is not clear whether it is an unbiased estimate $P(\mathbf{M}(p)|\mathbf{M}(q))$ for finite N. In fact, since $\mathbf{S}$ are biased estimates for $\mathbf{M}(p)$, it is reasonable to suspect that $P(\mathbf{S}|\mathbf{M}(q))$ is biased as well.

Given the central limit result for $\mathbf{S}$, a precise definition for the asymptotic 95% confidence ellipsoid can be made. Since, under the null hypothesis, $X_1,\ldots,X_N$ are q-distributed random variables, the confidence ellipsoid for $\mathbf{S}$ will have the form

$$\mathscr{E}(\mathbf{M}(q)) = \left\{\mathbf{a} : |D^{-1}(\mathbf{a} - \mathbf{M}(q))| \leqslant \frac{\zeta_{.95}}{\sqrt{N}}\right\}, \tag{A.11}$$

where $D = \sqrt{C}$, C is the covariance matrix defined in Proposition A.3, and $\zeta_{0.95}$ satisfies $\int_{B(0,\zeta_{0.95})}(2\pi)^{-K/2}\exp[-|\mathbf{x}|^2/2]\,\mathrm{d}\mathbf{x} = 0.95$. This confidence ellipsoid becomes asymptotically exact as $N \to \infty$. □

A.1. Comment on non-Gaussianity calculation

For the non-Gaussianity calculations, the statistics which are used do not depend upon the random mean S_1, but on the known mean s_1. These statistics, $\tilde{\mathbf{S}} = (S_3, S_4)$, satisfy a central limit theorem with the 2×2 covariance matrix given by the appropriate submatrix of Σ, in Eq. (A.7). In this definition of Σ, the PDF q is a Gaussian density with mean s_1 and variance s_2.

Appendix B. Details of central limit methodology

This section contains the statement and proof of a central limit theorem for $P(\mathbf{S}|\mathbf{M}(q))$, which is used for the central limit methodology in Section 5. The theorem follows from the central limit theorem proved in Proposition A.3 and Slutsky's theorem, Theorem A.2.

Theorem B.4. *The random variable $\sqrt{N}(P(\mathbf{S}|\mathbf{M}(q)) - P(\mathbf{M}(p)|\mathbf{M}(q)))$ converges in distribution to a Gaussian distribution with mean zero and variance $\sigma^2 = \mathbf{a}\ C\mathbf{a}^T$, where C is defined as in Proposition* A.3 *with p in place of q and*

$$\mathbf{a} = \boldsymbol{\theta} - (3M_2(p)\theta_3 + 4M_3(p)\theta_4)\mathbf{e}_1. \tag{B.1}$$

Proof. The proof is carried out for the case when $K = 4$. The proof of the $K = 2$ case is similar and simpler. Let $Z_N = \sqrt{N}(P(\mathbf{S}|\mathbf{M}(q)) - P(\mathbf{M}(p)|\mathbf{M}(q)))$. With a simple regrouping of the terms, it follows from Eq. (A.5) that

$$\begin{aligned} Z_N &= -\sqrt{N}(\log(Z(\boldsymbol{\theta}^N, S_1)) + \log(Z(\boldsymbol{\theta}, v))) + \sqrt{N}(\boldsymbol{\theta}^N \cdot \mathbf{S} - \boldsymbol{\theta} \cdot \mathbf{M}(p)) \\ &= \boldsymbol{\theta}^N \cdot \sqrt{N}(\mathbf{S} - \mathbf{M}(p)) - \sqrt{N}[\log(Z(\boldsymbol{\theta}^N, S_1)) - \log(Z(\boldsymbol{\theta}, v)) - \mathbf{M}(p) \cdot (\boldsymbol{\theta}^N - \boldsymbol{\theta})]. \end{aligned} \tag{B.2}$$

It follows from Slutsky's Theorem and Proposition A.3 that the first term, $\boldsymbol{\theta}^N \cdot \sqrt{N}(\mathbf{S} - \mathbf{M}(p))$, converges to a Gaussian distribution with mean zero and variance $\boldsymbol{\theta} \boldsymbol{C} \boldsymbol{\theta}^{\mathrm{T}}$. To determine the asymptotics of the second term, $Z(\boldsymbol{\theta}^N, S_1)$ is expanded about v and then about $\boldsymbol{\theta}$. With the change of variables

$$\gamma_j^N = \sum_{k=j}^{4} \binom{k}{j} \theta_k^N (v - S_1)^{k-j},$$

$\log(Z(\boldsymbol{\theta}^N, S_1))$ can be rewritten as

$$\begin{aligned} \log(Z(\boldsymbol{\theta}^N, S_1)) &= \log\left(\int \exp[\theta_1^N x + \sum_{k=2}^{4} \theta_k^N (x - S_1)^k] q(x)\, \mathrm{d}x\right) \\ &= \sum_{k=2}^{4} \theta_k^N ((v - S_1)^k - vk(v - S_1)^{k-1}) + \log\left(\int \exp[\gamma_1^N x + \sum_{j=2}^{4} \gamma_j^N (x - v)^j] q(x)\, \mathrm{d}x\right) \\ &= B_N + \log(Z(\boldsymbol{\gamma}^N, v)), \end{aligned} \tag{B.3}$$

where B_N denotes the first term in the previous display. Expanding $\log(Z(\boldsymbol{\gamma}^N, v))$ about $\boldsymbol{\theta}$, it follows that:

$$\begin{aligned} \log(Z(\boldsymbol{\gamma}^N, v)) &= \log(Z(\boldsymbol{\theta}, v)) + \mathbf{M}(p) \cdot (\boldsymbol{\gamma}^N - \boldsymbol{\theta}) + R_N \\ &= \log(Z(\boldsymbol{\theta}, v)) + \mathbf{M}(p) \cdot (\boldsymbol{\theta}^N - \boldsymbol{\theta}) + \mathbf{M}(p) \cdot \boldsymbol{\eta}^N + R_N, \end{aligned} \tag{B.4}$$

where

$$\eta_j^N = \sum_{k=j+1}^{4} \binom{k}{j} \theta_k^N (v - S_1)^{k-j} \quad \text{for } j = 1, 2, 3 \text{ and } \eta_4^N = \theta_4^N$$

and R_N is a remainder term of order $|\boldsymbol{\gamma}^N - \boldsymbol{\theta}|^2$. It follows from Eqs. (B.3) and (B.4) that the second term in Eq. (B.2) simplifies to

$$\sqrt{N}(B_N + \mathbf{M}(p) \cdot \boldsymbol{\eta}^N + R_N). \tag{B.5}$$

The central limit theorem implies that $\sqrt{N}(S_1 - v)$ converges to a Gaussian random variable. It follows from Slutsky's Theorem that $\sqrt{N}(S_1 - v)^k$ converges to zero for $k > 1$. Since

$$\sqrt{N}\mathbf{M}(p) \cdot \boldsymbol{\eta}^N = -\mathbf{M}(p) \cdot \begin{pmatrix} 2\theta_2^N \\ 3\theta_3^N \\ 4\theta_4^N \\ 0 \end{pmatrix} \sqrt{N}(S_1 - v) + \mathrm{O}(\sqrt{N}(S_1 - v)^2)$$

and

$$\sqrt{N}B_N = 2v\theta_2^N \sqrt{N}(S_1 - v) + \mathrm{O}(\sqrt{N}(S_1 - v)^2),$$

it follows that

$$\sqrt{N}(B_N + \mathbf{M}(p)\cdot\boldsymbol{\eta}^N) = -(3M_2(p)\theta_3^N + 4M_3(p)\theta_4^N)\sqrt{N}(S_1 - v) + \mathrm{O}(\sqrt{N}(S_1 - v)^2). \tag{B.6}$$

Substituting Eqs. (B.5) and (B.6) into Eq. (B.2), it follows that:

$$\begin{aligned} Z_N &= \boldsymbol{\theta}^N \cdot (\sqrt{N}(\mathbf{S} - \mathbf{M}(p))) - (3M_2(p)\theta_3^N + 4M_3(p)\theta_4^N)\sqrt{N}(S_1 - v) + \mathrm{O}(\sqrt{N}(S_1 - v)^2) + \sqrt{N}R_N \\ &= \mathbf{a}^N \cdot \sqrt{N}(\mathbf{S} - \mathbf{M}(p)) + \mathrm{O}(\sqrt{N}(S_1 - v)^2) + \sqrt{N}R_N, \end{aligned}$$

where $\mathbf{a}^N$ is the defined as in Eq. (B.1), but with $\boldsymbol{\theta}^N$ in place of $\boldsymbol{\theta}$. It follows from Slutsky's Theorem and Proposition A.3 that $\mathbf{a}^N \cdot \sqrt{N}(\mathbf{S} - \mathbf{M}(p)) + \mathrm{O}(\sqrt{N}(S_1 - v)^2)$ converges to a Gaussian distribution with mean zero and variance $\sigma^2 = \mathbf{a}C\mathbf{a}^{\mathrm{T}}$.

It remains to be shown that $\sqrt{N}R_N$ converges to zero in distribution. This fact follows from the well-known fact that if $\boldsymbol{\alpha}^N$ is a maximum likelihood estimator (MLE) of $\boldsymbol{\alpha}$, then $\sqrt{N}(\boldsymbol{\alpha}^N - \boldsymbol{\alpha})$ converges to a Gaussian distribution with mean zero and covariance given by the Fisher information matrix. To see that γ^N is the MLE of $\boldsymbol{\theta}$, first consider the set of multipliers $\boldsymbol{\alpha}^N$ such that

$$p_K(x) = Z(\boldsymbol{\alpha}^N)^{-1} \exp\left[\sum_{k=1}^{K} \alpha_k^N x^k\right] q(x).$$

The $\boldsymbol{\alpha}^N$ are chosen such that the uncentered moments of p_K equal the uncentered sample moments. In [30], it is shown that such an estimator is the MLE for the true parameters $\boldsymbol{\alpha}$. Since $\boldsymbol{\gamma}^N$ can be computed from $\boldsymbol{\alpha}^N$ via the linear transformation

$$\gamma_j^N = \sum_{k=j}^{4} \binom{k}{j} \alpha_k^N v^{k-j},$$

it follows that $\boldsymbol{\gamma}^N$ is the MLE of $\boldsymbol{\theta}$. It therefore follows from the fact that R_N is of order $|\boldsymbol{\gamma}^N - \boldsymbol{\theta}|^2$, that $\sqrt{N}R_N$ converges to zero in distribution. This completes the proof of the theorem. □

B.1. Comment on the central limit theorem result for non-Gaussianity

For the non-Gaussianity calculations, the statistics which are used do not depend upon the random mean S_1, but on the known mean s_1. This leads to a much simpler calculation of the variance than in Theorem B.4. Here, $\sigma^2 = \boldsymbol{\theta}C\boldsymbol{\theta}^{\mathrm{T}}$, where $\boldsymbol{\theta}$ are the Lagrange multipliers for the third and fourth moment constraints, and the matrix C is the 2×2 covariance matrix given by the appropriate submatrix of Σ, in Eq. (A.7) (See the comment at the end of Appendix A).

References

[1] G. Carnevale, G. Holloway, Information decay and the predictability of turbulent flows, J. Fluid Mech. 116 (1982) 115–121.

[2] T. Schneider, S. Griffies, A conceptual framework for predictability studies, J. Climate 12 (1999) 3133–3155.

[3] M. Roulston, L. Smith, Evaluating probabilistic forecasts using information theory, Mon. Weather Rev. 130 (2002) 1653–1660.

[4] L.-Y. Leung, G. North, Information theory and climate prediction, J. Climate 3 (1990) 5–14.

[5] R. Kleeman, Measuring dynamical prediction utility using relative entropy, J. Atmos. Sci. 59 (2002) 2057–2072.

[6] R. Kleeman, A. Majda, I. Timofeyev, Quantifying predictability in a model with statistical features of the atmosphere, Proc. Natl. Acad. Sci. USA 99 (2002) 15291–15296.

[7] A. Majda, I. Timofeyev, Remarkable statistical behavior for truncated Burgers–Hopf dynamics, Proc. Natl. Acad. Sci. USA 97 (23) (2000) 12413–12417.

[8] A. Majda, I. Timofeyev, Statistical mechanics for truncations of the Burgers–Hopf equation: a model for intrinsic stochastic behavior with scaling, Milan J. Math. 70 (1) (2002) 39–96.

[9] R. Abramov, G. Kovačič, A. Majda, Hamiltonian structure and statistically relevant conserved quantities for the truncated Burgers–Hopf equation, Comm. Pure Appl. Math. 56 (2003) 1–46.

[10] A. Majda, R. Kleeman, D. Cai, A framework for predictability through relative entropy, Meth. Appl. Anal. 9 (2002) 425–444.

[11] L. Mead, N. Papanicolaou, Maximum entropy in the problem of moments, J. Math. Phys. 25 (1984) 2404–2417.

[12] R. Abramov, A. Majda, Quantifying uncertainty for non-Gaussian ensembles in complex systems, SIAM J. Sci. Comp. 26 (2004) 411–447.

[13] D. Cai, K. Haven, A. Majda, Quantifying predictability in a simple model with complex features, Stoch. Dynam. 4 (2004) 547–569.

[14] R. Kleeman, A. Majda, Predictability in a model of geophysical turbulence, J. Atmos. Sci. (in press).

[15] E. Lorenz, K. Emanuel, Optimal sites for supplementary weather observations, J. Atmos. Sci. 55 (1998) 399–414.

[16] R. Abramov, A. Majda, R. Kleeman, Information theory and predictability for low frequency variability, J. Atmos. Sci. (in press).

[17] J. Anderson, W. Stern, Evaluating the potential predictive utility of ensemble forecasts, J. Climate 9 (1996) 260–269.

[18] Z. Toth, E. Kalnay, Ensemble forecasting at NMC: the generation of perturbations, Bull. Am. Meteorol. Soc. 74 (1993) 2317–2330.

[19] T. Palmer, Predicting uncertainty in forecasts of weather and climate, Rep. Prog. Phys. 63 (2000) 71–116.

[20] M. Ehrendorfer, J. Tribbia, Optimal prediction of forecast error covariances through singular vectors, J. Atmos. Sci. 54 (1997) 286–313.

[21] E. Kalnay, Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, New York, 2003.

[22] T. Palmer, F. Molteni, R. Mureau, R. Buizza, P. Chapelet, J. Tribbia, Ensemble predictionProceedings of the Validation of Models Over Europe, vol. 1, 1993, pp. 21–66.

[23] C. Reynolds, T. Palmer, Decaying singular vectors and their impact on analysis and forecast correction, J. Atmos. Sci. 55 (1998) 2576–2596.

[24] R. Buizza, T. Palmer, Impact of ensemble size on ensemble prediction, Mon. Weather Rev. 126 (1998) 2503–2518.

[25] E. Lorenz, Predictability: a problem partly solved, in: Proceedings of the Seminar on Predictability, ECMWF, Shinfield Park, Reading, England, 1996.

[26] T. Cover, J. Thomas, Elements of Information Theory, Wiley, New York, 1991.

[27] R. Blahut, Principles and Practice of Information Theory, Addison-Wesley, Boston, 1987.

[28] D. Williams, Weighing the Odds: A Course in Probability and Statistics, Cambridge University Press, New York, 2001.

[29] R. Abramov, A unified computational framework for the moment-constrained maximum entropy principle and its practical implementation, J. Comp. Phys. (submitted).

[30] R. Davidson, D. Solomon, Moment-type estimation in the exponential family, Commun. Stat. 3 (1974) 1101–1108.