Predictability in a model of geophysical turbulence

Richard Kleeman and Andrew J. Majda*

21st February 2004

Courant Institute of Mathematical Sciences, 251 Mercer Street, New York NY 10012.

Abstract

The nature of predictability is examined in a numerical model relevant to the mid-latitude atmosphere and oceans. The approach followed is novel and uses new theoretical tools from information theory. Particular attention is paid here to the practical application of these methods to the problem of ensemble prediction in dynamical systems with state spaces of high dimensionality. In this case typically only an estimate of the prediction probability distribution function is available either at coarse resolution or via (imperfect) knowledge of low order moments. A methodology for estimating the information loss implied by this limited knowledge is introduced and applied to the practical problem of measuring prediction information content in a model able to generate geophysical turbulence. The application studied here generates such turbulence through the mechanism of baroclinic instability via an imposed and constant mean vertical shear. In traditional studies in this area considerable attention has been paid to variations in ensemble spread as the major determinant of how predictability may change as prediction initial conditions vary. Our analysis reveals that such a scenario neglects the important role of the so-called ensemble signal which is related to the difference in the first moments of the prediction and climatological distributions. We find in fact that this quantity is often a strong control over variations in predictability. An initial investigation of the role of non-Gaussian effects shows that for the univariate large scale barotropic case they are only of minor importance to variations in predictability. This is despite the fact that for long prediction times, non-Gaussian effects can be quite noticeable in the prediction distribution.

*Corresponding Author: Dr. Richard Kleeman, Courant Institute, 251 Mercer Street, New York NY 10012, email: kleeman@cims.nyu.edu

1 Introduction

The problem of weather prediction has a long and interesting history both from a theoretical and practical perspective. Lorenz [17] was amongst the first to recognize the extreme sensitivity of such predictions to small variations in the specification of initial conditions. In a series of papers in the succeeding decades Lorenz essentially initiated some of the present considerable interest in chaotic dynamical systems (e.g. Ruelle and Takens [26] and Grassberger [11]). Later a theoretical framework for statistical prediction involving probability distribution functions ($p.d.f.'s$) was proposed by a number of authors including for example Epstein [9] and Leith [16]. This approach has been particularly useful from a pedagogical viewpoint. From a practical perspective the problem of how to implement the program of statistical prediction received much attention in the past two decades (eg. Murphy [22], Toth and Kalnay [32], [33], Palmer et. al. [23], Molteni et. al. [21], Houtekamer et. al. [13], Ehrendorfer and Tribbia [8] and Buizza and Palmer [5]). The basic difficulty here is that generally only a relatively small ensemble estimate of the prediction $p.d.f.$ is practically available in what is a high dimensional dynamical system and in a situation where higher order moments of the $p.d.f.$ may contribute significantly (see Kleeman [14]).

Recently there has been a renewed theoretical interest in predictability particularly from an information theoretical perspective (Carnevale and Holloway [6], Schneider and Griffies [30], Kleeman [14] (henceforth referred to as K1) and Roulston and Smith [25]). Such an approach is particularly attractive as it allows us to define transparent measures of the information content of predictions which have attractive theoretical properties. In particular K1 has recently advocated the use of the relative entropy of the prediction and climatological $p.d.f.'s$ as a measure of the utility of a particular statistical prediction. Conceptually before a prediction is made knowledge of a particular dynamical system is given by the climatological or equilibrium $p.d.f.$. Once the prediction $p.d.f.$ is available the informational inefficiency of ignoring this and using the prior knowledge is given by the relative entropy of the two distributions. Not surprisingly such a measure has also received considerable attention in the statistics literature (Bernardo and Smith [3] gives a good overview and we will refer to this text as BS henceforth) where it forms a part of the foundations of Bayesian theory. Here it is also referred to as a utility measure. In generic non-rigorous terms we may formulate it as

$$D(p(x), q(x)) \equiv \int p(x) \ln \left(\frac{p(x)}{q(x)} \right) dx \tag{1}$$

where $p(x)$ is the prediction $p.d.f.$ while $q(x)$ is the equilibrium $p.d.f.$ which can be considered to be periodic in time.

In general it is often a reasonable approximation to consider the time evolution of random variables in geophysical applications to be approximately Markovian. Such a property will hold if the entire nature of the $p.d.f.$ at time step $t+1$ can be derived from that at time t. Clearly the numerical formulation of most geophysical problems in time stepping form often[1] ensures that their numerical approximation satisfies such a property. If the Markovian property holds then the relative entropy satisfies three particularly attractive properties:

1. $D(p(t), q(t)) \geq 0 \quad \forall t$ **Positivity**
2. $D(p(t_1), q(t_1)) \geq D(p(t_2), q(t_2)) \quad t_2 > t_1$ **Temporal Monotonicity**
3. $D(p(\alpha), q(\alpha)) = D(P(F(\alpha)), Q(P(F(\alpha)))$ **Invariance**

where $F : \Psi \rightarrow \Phi$ is a general non-linear transformation of state-space variables with non-zero Jacobian. Rigorous demonstrations of the first two properties can be found in Cover and Thomas [7] while the latter is shown in BS (page 158) and Majda et. al. [18] (henceforth referred to as MKC). It is worth emphasizing that other entropic measures do not satisfy any of these properties in general.

An interesting aspect of the use of this measure is the connection to non-equilibrium statistical mechanics. The second property above can be interpreted as a generalized second law of thermodynamics for Markov processes. In molecular statistical dynamical systems where this law was first proposed over a century ago by Boltzmann [4], the equilibrium *p.d.f.* is actually uniform on an energy hypersphere and in this case if we assume energy conservation then the relative entropy reduces to minus the absolute (standard) entropy (see equation 1, above) and so the usual formulation of the second law is recovered. Evidently in many problems of practical geophysical interest, the equilibrium *p.d.f.* is far from uniform and in this case the relative entropy emerges as a particularly natural measure. In terms of the analogy with statistical mechanics, property 2 above shows that it is *the degree of disequilibrium* of the system at prediction time that measures the usefulness of the prediction.

In K1 a range of stochastic models were examined which have direct relevance to the problem of climate prediction (the stochastic forcing represents the atmospheric time scales here) and it was found that the first moment of the prediction *p.d.f.* was often the major control on variations in the utility of predictions with initial conditions. In fact for stochastic differential equations with constant coefficients it may be shown almost trivially that for deterministic initial conditions the only control on utility variation is the first moment. This situation contrasts strongly with that normally assumed in atmospheric prediction where it often assumed that it is the *p.d.f.* spread (the second moment) that exercises such a control. A natural question then to ask is whether such a common assumption is justified within the general formulation of predictability that we have proposed. A first examination of this question was undertaken by Kleeman et. al. [15] using a particularly idealized model of the atmosphere and ocean known as the truncated Burgers model (see Majda and Timofeyev [19]). In this case the *p.d.f.'s* are often close to Gaussian which allows us to approximately calculate utility through exact formulae (see below for a more rigorous

[1] Of course if some aspect of the problem depends on values of the state-space variables from several time steps back in time this will not hold. Such situations are, however, not very common.

and revealing discussion of such an approximation). It was found that the first moment again plays a very important role in controlling utility variations. Motivated by these results and the desire to introduce general techniques to deal with non-Gaussian $p.d.f.'s$ and finite ensembles we re-examine these ideas here in a model of geostrophic turbulence which has many of the physical features of the mid-latitude atmosphere and ocean.

As alluded to above, a major problem in statistical prediction concerns calculation of the prediction $p.d.f.$ and its evolution. In practical situations this function is estimated by the Monte Carlo technique of ensemble prediction. Here one draws initial conditions according to some estimate of the initial condition $p.d.f.$ and calculates many trajectories. As the dimension of the state space increases however the estimation of the $p.d.f.$ becomes more and more difficult, a situation sometimes referred to as "the curse of dimensionality". Alternatively one may attempt to approximate the $p.d.f.$ through calculation of sample moments (Mead and Papanicolaou [20], MKC and Abramov and Majda [1]). Again one faces a limit as the estimates of moments become less and less reliable as the order of the moments considered increases. The perspective we shall adopt here in response to this problem is derived from information theory: An ensemble estimate evidently implies a reduction in the amount of information known about the prediction and as we shall see in the next section it is possible to quantify this loss of information rather precisely. Thus our philosophy is that ensemble prediction implies an information loss over the ideal of $p.d.f.$ prediction (which has been studied previously). The nature of this information loss is rather interesting and evidently of significant practical interest. We only begin our exploration of its consequences here.

The remainder of this contribution is structured as follows: In section 2 we develop the tools for calculating the information content of an ensemble prediction. In section 3 the dynamical model to be studied is introduced, justified and explored. In section 4 the methodology of Section 2 is applied to the dynamical model and the question of what controls variations in utility addressed. Section 5 contains a discussion and summary of the results.

2 Information loss due to ensemble estimation

As mentioned previously it is usually impossible to calculate the full evolution of the prediction $p.d.f.$ since the state spaces of dynamical systems of practical interest are often of very high dimensionality. In addition the structure of the $p.d.f.$ can sometimes become highly non-Gaussian and again as a consequence difficult to estimate. As an example, K1 showed that the standard Lorenz 3 mode model has this property. Usually only a Monte Carlo estimation known as an ensemble is available and often the size of this sample of the prediction $p.d.f.$ is considerably smaller than the state-space dimension. Various selective sampling techniques (singular and breeding vectors) are often deployed in an attempt to circumvent this problem (see, for example, Palmer and Tibaldi [24] and Toth and Kalnay [33]). Here we adopt a different approach: An ensemble estimate implies that the full information of the prediction *p.d.f.* is fundamentally unavailable to us. Indeed it may be shown rather easily (see below) that for realistic situations there is considerable loss of information implied for any ensemble that is within practical reach. We present here two methods for calculating such an information loss. The first relies on a coarse-graining of a relevant subspace of state-space while the second calculates the information content of successively higher order moments and the information loss implied by the less than perfect knowledge of these quantities due to their sample (i.e. imperfect) estimation from our ensemble. As we shall see there are two sources of information reduction. The first is due to the coarse-graining or the moment truncation strategy adopted while the second is due to sampling error with respect to this chosen coarse graining or moment truncation strategy.

2.1 Coarse-grained ensemble estimation

In geophysical applications it is often possible to explain much of the variability of many dynamically relevant variables in a very high dimensional dynamical system with relatively few modes. These modes often tend to be large scale spatially and of low frequency. This situation holds for the model we shall consider later. Providing that the number of these modes is not large we can obtain useful estimates from ensembles of their information content albeit at fairly coarse resolution.

Let us suppose that this reduced space has dimension n and that we have a complete partitioning of this space by subsets X_i with $i = 1, ..., m$. In general one would expect that $m \gg n$ in order that there be adequate resolution of each dimension in our coarse-graining. Given such a partitioning then an ensemble implies a frequency count f_i associated with every element X_i. Providing that $m \gg n$ then many of the f_i may be of significant size. As a concrete illustration let us suppose we are interested in quartile information for each dimension then we require ensembles of size at least 4^n for there to be many f_i of significant size. For ensembles of size $10^3 - 10^4$ (the usual practical limit) this implies that approximately $n < 7$ at least for quartile resolution. Higher resolution evidently requires larger ensembles.

Consider now the prediction *p.d.f.* p on this reduced state-space. If we integrate over each partition element X_i we obtain the coarse-grained discrete probability vector element p_i. Evidently we could estimate such a vector using the f_i. Consider now the conditional probability $P(\mathbf{f}|\mathbf{p})$ that we observe f_i given that p_i holds. It follows from elementary probability theory that

$$P(\mathbf{f}|\mathbf{p}) = \frac{1}{Z(\mathbf{f})} \prod_{i=1}^{m} p_i^{f_i}$$

$$Z(\mathbf{f}) = \frac{\prod_{i=1}^{m} (f_i)!}{(N + m - 1)!}$$

$$N = \sum_{i=1}^{m} f_i$$

Now Bayes theorem (see BS) gives

$$P(\mathbf{p}|\mathbf{f}) \propto P(\mathbf{f}|\mathbf{p})P_{pr}(\mathbf{p})$$

where $P_{pr}(\mathbf{p})$ is the prior[2] probability that a particular set of p_i occurs. Without any evidence of what values p_i take it is reasonable to take this prior probability as uniform i.e. in the absence of evidence there is no reason to expect that any one set of p_i is any more likely than any other. With this assumption we obtain

$$P(\mathbf{p}|\mathbf{f}) = \frac{C}{Z(\mathbf{f})}\prod_{i=1}^{m} p_i^{f_i} \equiv \Phi_{\mathbf{f}+1}(\mathbf{p})$$

where Φ is the so-called Dirichlet distribution. The first moment of this distribution then gives the expected p_i given the observed f_i and may be shown easily to be given by

$$\langle p_i \rangle = \frac{f_i + 1}{N + m}$$

Now that we have calculated the distribution of the p_i we can calculate the expected information loss in assuming that $p_i = \langle p_i \rangle$. This is clearly

$$EL(\mathbf{f}) = \int \Phi_{\mathbf{f}+1}(\mathbf{p}) D(\mathbf{p}, \langle \mathbf{p} \rangle) d\mathbf{p}$$

where $D(\mathbf{p}, \langle \mathbf{p} \rangle)$ is the relative entropy of the coarse-grained $p.d.f.'s$ $\mathbf{p}$ and $\langle \mathbf{p} \rangle$. Using known analytical expression for moments of the Dirichlet distribution and the expected values of their logarithms it is possible to evaluate EL analytically with the result:

$$EL(\mathbf{f}) = \sum_{i=1}^{m} \langle p_i \rangle \left(\psi(f_i + 2) - \psi(n + m + 1) - \ln \langle p_i \rangle\right) \tag{2}$$

where ψ is the digamma function (Abramovitz and Stegun [**?**] p258).

It may also be shown relatively straightforwardly that the expected information loss is minimized by choosing $\langle p_i \rangle$ as our estimator of the coarse-grained $p.d.f.$ p_i. The relative entropy of the coarse grained optimal prediction and climatological $p.d.f.'s$ $\langle p \rangle$ and $\langle q \rangle$ may be easily evaluated:

$$D(\langle p \rangle, \langle q \rangle) = \sum_{i=1}^{m} \langle p_i \rangle \ln\left(\frac{\langle p_i \rangle}{\langle q_i \rangle}\right) \tag{3}$$

In general, in most practical contexts the loss of information due to sampling of the climatological distribution is considerably smaller than that due to the prediction ensemble sampling since it is much larger. An intuitively appealing definition then of the utility of an ensemble prediction is therefore:

$$EU(\langle p \rangle, \langle q \rangle) \equiv max\left\{D(\langle p \rangle, \langle q \rangle) - EL(\langle p \rangle), 0.0\right\} \tag{4}$$

i.e. one reduces the coarse grained information content by the information loss due to the prediction distribution sampling.

[2]In other words prior to the ensemble observation of the frequencies f_i

2.2 Moment truncation ensemble estimation

In general for finite samples, the accuracy of moment estimators declines as the order of that moment increases. Thus an alternative to the coarse-graining strategy of the previous section is to retain only the information content of the moments which are "accurately" estimated by an ensemble of a given size. As a concrete illustration one might expect intuitively that for ensembles of order 10^2 only the first and second moments may be accurately estimated and so in calculating the information content of our prediction ensemble, we should only retain that due to these particular moments. Evidently we need some concrete estimate of the information loss due to the inaccuracy of our moment estimation in the same way as we needed a loss function for a particular partitioning of state-space in the previous section. Further we require an estimate for the information content due to the retained lower order moments. The most natural estimation procedure for the *p.d.f.* which has moments given by the truncated set is the very well known maximum entropy approach (see e.g., BS p207, Mead and Papanicolaou [20], MKC and Abramov and Majda [1]). This produces the *p.d.f.* which exhibits the maximal ignorance with respect to the neglected higher order moments. In other words the distribution which ignores their information content. Rather nicely when one neglects the information content of the third and higher moments one obtains the standard Gaussian distribution specified by the retained first and second moments and so in this case the expression for the relative entropy reduces to the analytical form already derived in K1. Thus this expression not only corresponds to the information content obtained by assuming all distributions are approximately Gaussian but also corresponds to the information content specified by the first two moments. In order to preserve readability of this contribution the rather technical discussion of how to carry out the program for moment truncation *p.d.f.* estimation and corresponding sampling information loss is detailed in the Appendix.

3 Quasigeostrophic turbulence: Model and basic results.

The mid-latitude dynamical system which underlies both the atmosphere and ocean has been extensively studied in the past few decades (see Salmon [29] for an excellent overview). Central to these studies has been the so-called quasigeostrophic approximation of the primitive equations. This holds, crudely speaking, if the Rossby number is significantly less than unity and the Coriolis parameter does not vary greatly. Physically the approximation has the effect of filtering gravity waves and confining attention to low frequency variability which is close to geostrophic balance. Many of the broad features of mid-latitude variability which result from baroclinic and barotropic instability are well captured by models incorporating such an approximation.

Our aim in this contribution is to study the nature of predictability of as simple a system as possible which still retains the dominant physical instability mechanisms of the mid-latitudes. The intention is to ensure that the main processes underlying turbulence generation in this region are retained in as simple a form as possible. This approach is motivated philosophically by the expectation that the basic predictability properties of the mid-latitude dynamical system should follow from the nature of the turbulence there.

The simplest model meeting the above criteria is a two level quasigeostrophic configuration which is externally forced by a mean vertical shear simulating the effect of differential meridional radiative forcing. We selected a two-layer version of the the particular model of Smith et. al. [31] since the properties and nature of its turbulent cascade have received extensive discussion in the literature (see Salmon [29] Chapter 6 and references cited therein as well as the comprehensive discussion and citations in Smith et. al. [31]).

The governing equations are prognostic in the potential vorticity q of the flow and for a two level model (Salmon [29] page 111) with surface Ekman damping; orography and a constant mean vertical shear $US \equiv U_1 - U_2$ may be written as

$$\frac{\partial q_i}{\partial t} + J(\psi_i + \overline{\psi_i}, q_i + \overline{q_i} + R(i)h_b) = -\kappa R(i)\nabla^2\psi_i + F_{hyp} \quad (5)$$

where ψ_i is the stream function; κ the Ekman damping coefficient; h_b the orographic height function and

$$R(1) = 0$$
$$R(2) = 1$$

The following relation holds between the potential vorticity and the stream function:

$$q_1 + \overline{q_1} = \nabla^2\psi_1 + \beta y + S\left((\psi_2 - \psi_1) + USy\right)$$
$$q_2 + \overline{q_2} = \nabla^2\psi_2 + \beta y + S\left((\psi_1 - \psi_2) - USy\right)$$

with $S = \frac{4f_0^2}{H^2N_m^2}$ (H is the troposphere height while N_m is the Brunt-Vaisalla frequency at the model midpoint vertically). The mean stream function is given by

$$\overline{\psi}_i = -U_i y$$

in order to assure that the background state is geostrophic. Finally the term F_{hyp} represents a hyperviscosity which in spectral space (the method used to solve the model) acts as a damping predominantly on the largest wave-numbers of the model. This term is used to ensure numerical stability and simulates the sink of energy at the smallest scales of the model (see Smith et. al. [31] for further discussion on the precise formulation). Non-dimensional parameter choices used in the numerical experiments below are detailed in Table 1. The symbols have the following meaning: f_0 is the Coriolis parameter about which the beta-plane used is constructed; $2\pi L$ is the domain size in both directions; g' is the mean reduced gravity for the two layer configuration; β is the meridional gradient of the Coriolis parameter while U_o is the horizontal velocity scale.

For our initial numerical experiments reported here we chose to use a doubly periodic domain and considered the effects of orography in an idealized

Table 1: Parameter values for the numerical experiments

Experiment	$F = \frac{f_0^2 L^2}{4\pi^2 g' H}$	$beta = \frac{\beta L^2}{2\pi U_0}$	Topography
Control	400.0	2.0	None
Large Rossby Radius	4.0	2.0	None
Orographic Forcing	400.0	2.0	Equation (7)

fashion (see below). Obviously these two configuration choices may affect our conclusions and more general numerical results will be reported in a future contribution. Given the numerically demanding ensemble experiments reported in the next section we chose to use a reasonably coarse horizontal resolution and retained 15 wave-numbers in the zonal and meridional directions. We tested the sensitivity of equilibrium behavior and timescale to a doubling of resolution in both directions and noted little qualitative change in model behavior.

When the model described by equations (5) with the parameters of Table 1 is integrated from arbitrary initial conditions the equilibration process is controlled primarily by the turbulent cascade rather than by Ekman spin-down. When scaling is chosen appropriate to atmospheric conditions, the timescale involved is order weeks rather than days (the Ekman timescale). The process by which the equilibrium turbulent cascade is maintained was studied by Salmon [27], [28] and Held and Larichev [12] and is displayed in Figure 12 of Salmon [29]: Energy injected by the large-scale (constant) mean shear into the baroclinic component of the model cascades via the non-linearities of the model to smaller scales until it reaches the baroclinic Rossby radius. At this scale transfer to the barotropic component of the model is possible. Barotropic energy at the conversion horizontal scale cascades primarily to the larger scales where it is removed by the Ekman dissipation. Energy also cascades in both the barotropic and baroclinic modes to scales smaller than the conversion scale where it is removed by the hyperviscosity term of the model. When equilibrium is achieved most energy occurs in the large-scale barotropic modes. This behavior is depicted in Figure 1 which depicts the barotropic energy spectrum of the equilibrium (a large time average is used) as well as a typical snapshot from both vertical levels. Parameters for this model run were the control set from Table 1.

An important consequence of the equilibrium state of the turbulence from the viewpoint of theoretical predictability studies is that relatively few large-scale barotropic modes are required to explain a considerable amount of important variability within the model. This was confirmed by performing a linear regression at each point of the domain between local stream function and the first two non-constant complex barotropic spectral modes for an extended time period during equilibrium. With respect to the two dimensional Fourier decomposition, the complex modes used have wavenumber vectors $(1, 0)$ and (0,1) . Given the complex nature of the modes obviously four degrees of freedom are involved. The explained variance in the control run was approximately constant at around 95% while with the idealized large scale orography considered (see below) it was slightly lower at around 92%. This situation changed noticeably when a large Rossby radius was used ($F = 4.0$). The barotropic spectrum then was not as peaked as in the control run and also significantly more baroclinic. In this case the large scale barotropic modes accounted for around 60% of the surface and 45% of the upper level stream function variance at any point. As we saw in the previous section, calculation of the information content of a prediction ensemble can be problematic when the dimension of the variables used is large because one then typically requires huge (and thus impractical) ensembles. The concentration of a large amount of variance into four modes in the present model enables us to partially circumvent this difficulty by concentrating on the utility associated with the large-scale barotropic modes. If we choose to retain information associated only with the first two moments and hence consider only Gaussian $p.d.f.'s$ then this difficulty may also be avoided and we do not need to restrict ourselves to a four mode state-space since in such a case we have analytical expressions for both the utility and the ensemble sampling information loss (see section 2.2). We followed the coarse graining strategy here since noticeable non-Gaussianicity was evident in prediction distributions. The Gaussian approach was followed in an earlier study (Kleeman et. al. [15] where this was not the case. Given the non-Gaussian distributions we will also investigate the degree to which the higher order moments contribute to the information content using the methodology described in the previous section.

The approach we follow here of concentrating on the predictability of the large scale barotropic modes of the flow will be extended in future studies to consider other important physical variables such as temperature which depends significantly on the baroclinic component of the flow. A related study by Abramov and Majda [1] with a simpler model has considered such variables.

4 Predictability Results

4.1 Experimental design

The motivation of the present study is to identify the nature of the variation in prediction utility with differing initial conditions. This variation is evidently of potentially great practical importance. In order to gain a representative view of such variations, we draw such initial conditions according to the climatological $p.d.f.$. This is done by performing an extended integration of the model after it has achieved equilibrium[3] and choosing the initial conditions at a sufficiently large equal time interval to ensure no correlation of the state variables (which we take to be the stream function) from one set of initial conditions to the next. At each initial condition set, an ensemble is generated by adding a small perturbation distributed according to a Gaussian with equal variance in all spectral components of the stream-function. The standard deviation of this distribution was chosen to be 0.1 dimensionless units. For comparison climatological

[3]This was identified by monitoring the total energy of the model and ensuring that it was quasi-steady in time.

standard deviation of the dominant spectral mode (the $(1,0)$ mode in spectral notation) is of order 30.0 units for the control and orographic experiments conducted here. Each ensemble member was integrated for 0.9 dimensionless time units which was roughly 60% of the way to equilibrium. A typical collection of ensemble member trajectories are depicted in Figure 2 which plots the evolution of the real part of the $(1,0)$ spectral mode at the upper model level. In order to explore the predictability concepts introduced in section 2 a rather large 1000 member ensemble was produced for each of 50 initial condition sets. In practical applications one is often restricted to smaller ensembles.

As a first exploration of the predictability of this physically relevant system[4] we chose to explore the sensitivity of predictability to variations in the following parameter F of the flow which is related to the square of the inverse of the Rossby radius of deformation:

$$F = \frac{f_0^2 L^2}{4\pi^2 g' H} \tag{6}$$

where H is the model vertical height while g' is the reduced gravity associated with the stratification which produces the mean vertical shear. We examined predictability at $F = 400.0$ and 4.0 which correspond very approximately to mid-latitude oceanic and atmospheric flow regimes respectively. Results were also calculated for the intermediate case of $F = 40.0$ but these were essentially qualitatively identical to the $F = 400.0$ case. In the case of the $F = 4.0$ setting, the climatological standard deviation of the dominant large scale modes were a factor of 20 reduced compared to the control case. We consequently reduced the ensemble initial condition perturbations by the same factor.

4.2 Coarse grained entropy

As we have seen, a large amount of variance in this model may be explained by the first two non-constant complex spectral modes. We chose therefore to partition a reduced four dimensional subspace. With a 1000 member ensemble we coarse grained each dimension into quartiles (with respect to the prediction ensemble) which implied that each partition box in the four dimensional space had $5-6$ members for the prediction ensemble. For the climatological ensemble we took a large number (10^5) of basically uncorrelated snapshots from an extended equilibrium integration of the model. Given the much larger climatological ensemble size, partitions often had large number of climatological ensemble members within them (order 1000) but also often very few depending on how far from equilibrium the prediction ensemble was. The relaxation to equilibrium of all the initial conditions (with $F = 400.$) is depicted in Figure 3 which shows the ensemble utility (see equation 4) as a function of time. Noteworthy is the significant fluctuation from one initial condition to another. The variations in utility are strongly related to those derived under the assumption

[4] A future publication will explore the role of boundary conditions, realistic orography and spherical geometry.

that all distributions are Gaussian (see subsection 2.2., Appendix A., K1 and Kleeman et. al. [15]). As was discussed in this latter case the utility R may be broken down into a term dependent on the first moments of the prediction *p.d.f.* which we term the "signal" and terms dependent on the second moments which we term the collectively the "dispersion":

$$\begin{aligned} R &= \frac{1}{2}\left[\ln\left(\frac{\det(\sigma_q^2)}{\det(\sigma_p^2)}\right) + tr\left(\sigma_p^2(\sigma_q^2)^{-1}\right) - n\right] \quad \textit{Dispersion} \\ &+ \quad \frac{1}{2}(\overrightarrow{\mu}^p - \overrightarrow{\mu}^q)^t(\sigma_q^2)^{-1}(\overrightarrow{\mu}^p - \overrightarrow{\mu}^q) \quad \textit{Signal} \end{aligned}$$

where σ^2 and $\overrightarrow{\mu}$ are the covariance matrices and mean vectors respectively and the sub/superscripts p and q refer to the prediction and climatological distributions. Variations in the coarse-grained utility here are generally strongly related to the signal term. This behavior is depicted in Figure 4 which shows the relationships at dimensionless times of 0.2 and 0.7 which might be considered short and medium range predictions in weather nomenclature. It is worth noting also that in this case the ensemble spread is generally not a good predictor of (coarse-grained) utility. The relaxation of *p.d.f.s* to equilibrium is rather interesting and is depicted in Figure 5. In general the climatological distribution is centered approximately on a sphere in the four dimensional phase space with the distribution with respect to the radius of this sphere being approximately Gaussian. Points on such a sphere represents equal energy configurations in our reduced state-space. Presumably this approximate energy conservation by the low wavenumber barotropic modes represents a balance between energy injection from the higher wavenumber (barotropic) modes and dissipation at large scales by the Ekman friction. Prediction distributions at short range are approximately Gaussian patches located at arbitrary points on or near this sphere (Figure 5c). As time increases (Figure 5d) this patch spreads (reasonably uniformly) around the sphere and hence when viewed univariately (Figures 5e) tends to become noticeably non-Gaussian due to the spherical geometry which is guiding its relaxation. We examine how important this non-Gaussian behaviour is to the entropy below.

4.3 Sensitivity to Rossby radius and orography

The parameter F (see equation (6) above) controls the square of the ratio of the domain size to the model Rossby radius so smaller values might be viewed (rather simplistically) as moving the model into a more "atmospheric" regime. One might expect a priori predictability properties to be sensitive to this parameter in view of results reported elsewhere by the author (K1). We therefore reduced F from 400 to 40 and then to 4 and repeated our experiments from the previous subsection. Interestingly, results obtained with $F = 40.0$ were virtually indistinguishable from the control case. In the case of $F = 4.0$ the equilibrium turbulent cascade implied a flow less dominated by the large-scale barotropic flow (see above). The coarse grained entropy of these modes was if anything even more dominated by the Gaussian signal (see Figure 6) at all time lags. Interestingly in this case there was a considerably less even spread in the entropy at all leads with a few very high utility cases and many low cases. A viewing of the high utility cases showed that they were consistently high for all prediction times.

Orography has often been cited as a factor in modifying climatological (and presumably prediction) $p.d.f.s$ (see e.g. Frederiksen [10]) so we tested the sensitivity of our results to it's inclusion. Following Smith (private communication) we introduced an idealized orography which was Gaussian (and isotropic) in spectral space with mean total wavenumber 3 and standard deviation of 2 wave-numbers and also randomized with respect to the phase of the spectral modes. In spectral space this may be written

$$\hat{h}_b(\overrightarrow{k}) \equiv \frac{\beta L}{f_0} H_L exp\left(-(|k| - k_{topo})^2/k_{dev}^2\right) exp(2\pi i \times ran(\overrightarrow{k})) \tag{7}$$

where k_{topo} and k_{dev} are the wavenumber amplitudes corresponding to wave numbers 3 and 2 respectively; $ran()$ is a random number in the interval $(0, 1)$. The parameter H_L is the height of the lower layer of the model. For parameter choices consistent with a mid-latitude global domain, the peak topographic height is rather large by terrestrial standards at around the height of the lower layer.

Predictability results were somewhat different from the control case (Figure 7): For short ranges the signal was still the dominant factor in utility variation however at a (subjectively) reduced rate. For longer range predictions, signal and dispersion appeared equally important to (coarse grained) utility variation. Interestingly the dispersion showed a "triangular" relationship to coarse grained utility with high dispersion being very often associated with high utility whereas low dispersion was associated with both low and high utility situations. Such a relationship has been often reported in weather predictions studies of the relationship between skill and spread (the latter being closely related to the dispersion studied here).

4.4 Importance of higher moments to predictability

In section 4.2 above we noted that for "medium range" predictions the $p.d.f.$ often shows some non-Gaussianicity due to the manner in which equilibration takes place. Given this we decided to check the potential of such an effect to influence entropic measures. A detailed framework for studying the effects of higher moments has been proposed elsewhere by MKC and will be applied comprehensively to a model similar the one studied here in a future publication by the authors and coworkers (see also section 2.2 and the Appendix). The MKC approach provides a computationally feasible strategy for computing a series of information content bounds and has been applied to a simple pedagogical model in Abramov and Majda [1]. As a first step towards applying this technique in a quasi-geostrophic context we examined the effects of including the third and fourth moments on the *univariate* relative entropy which we define as the sum of all the relative entropies of the marginal distributions:

$$D_u(p,q) \equiv \sum_{i=1}^{N} D(p_i, q_i) \tag{8}$$

where p_i and q_i are the marginal distributions of the full $p.d.f.'s$ $p(x_1, x_2, x_3, \ldots, x_N)$ and $q(x_1, x_2, x_3, \ldots, x_N)$ with respect to the state variable x_i. Computationally the program detailed in Section 2.2 and the Appendix for computing entropic functionals is most straightforward for univariate distributions such as $p_i(x_i)$ since the unconstrained minimization (maximum entropy) algorithm detailed there works in a one dimensional space for which there is an extensive literature. We restricted our attention for this computation to the times for which the univariate distributions were noticeably non-Gaussian namely "medium range" predictions. Plotted in Figure 8 is D_u the first four Fourier modes discussed previously. The third and fourth moments of the full marginal distributions are used to calculate the p_i included in equation (8). What is shown is the relationship between the univariate entropy when these additional moments are included versus that when only the first two moments are included i.e. the Gaussian case for which there are analytical expressions. Figure 8a shows the control case with $F = 400$; 8b shows the case in which orography is included while 8c is the case with $F = 4.0$. Results are shown for the same time as that in Figure 4cd and 7cd. In general inclusion of the univariate third and fourth moments adds some information however the relationship with the Gaussian case is very strong suggesting that most of the information at least in the univariate case is coming from the first and second moments. This result is a little surprising given the non-Gaussian nature of the univariate distributions seen above however one must keep in mind that the prediction distributions means and variances are often very different from that of the climatology so the *relative* contribution of the higher moments may not be large. One must also stress that the cross (multivariate) terms which we have not included in this calculation (see MKC) may alter the preliminary conclusions presented here. One interesting effect is evident in the "atmospheric" $F = 4$ case (Figure 8c). Here the non-Gaussian contribution increases as the total utility (and Gaussian signal) increases suggesting that prediction distributions showing a large "difference" from the climatological distribution are the most non-Gaussian.

5 Summary and Discussion

A useful way of analyzing predictability in dynamical systems is through examination of the relaxation of prediction (probability) distributions towards a quasi-stationary equilibrium distribution often called the climatological distribution. The degree of this disequilibrium may be measured rather precisely by the relative entropy of the two distributions. This functional corresponds with the informational inefficiency of assuming the climatological distribution when in fact the prediction distribution holds. From a Bayesian perspective it thus represents the additional information brought to the table through the prediction process: Here one identifies the prior distribution with climatology and the posterior with the prediction distribution. Given this background the relative entropy measures rather transparently the utility of the prediction process.

In addition, the relative entropy satisfies a number of elegant mathematical properties including perhaps most importantly invariance under non-degenerate nonlinear transformations of state variables. This latter property would appear almost mandatory for a measure of predictability in geophysical systems where such state variable transformations are common.

In practical situations the full multivariate prediction and climatological distributions are unavailable since their time integration rapidly becomes infeasible as the dimension of state space increases. Instead one normally relies on Monte Carlo or ensemble methods to sample such distributions. This process must involve some reduction in information and hence the utility of the statistical predictions made. In order to analyze this loss one needs to adopt a particular "coarse graining" frame of reference since point-wise probability density estimates are obviously impossible. Two natural possibilities for this are the geometric partitioning of state space and the retention of a finite number of the low order moments from the distributions.

Two forms of information loss are associated with these coarse graining frames. First the very act of coarse graining implies a discarding of information associated with the fine scales or the neglected higher moments. Secondly the remaining coarse grained quantities are subject to sampling error which again implies information loss. These two forms of loss are evidently connected since as one refines the coarse graining one should expect the sampling error of the "finer" quantities to become larger. Such a trade-off in information loss could in fact be used to define an optimal coarse graining, a subject we will pursue further in a future publication.

We analyzed in detail here the two coarse graining strategies and derived expressions for the sampling information loss. The coarse grained relative entropy is easily calculated in the geometric partitioning case however in the moment truncation case one needs to define an appropriate distribution which has the retained moments. This can be done by assuming "maximal ignorance" with respect to the neglected higher order moments or equivalently assuming that no information is available on the neglected moments. This "maximum entropy" principle is widely known in statistical and physical applications and from a mathematical viewpoint defines an optimization problem with respect to possible distributions. A natural byproduct of this optimization process is the relative entropy. Very interestingly in the case that only the first and second moment information is retained then the distribution obtained is that which would have been derived under the assumption that all distributions were Gaussian. In such a case analytical expressions are possible for all relevant quantities and results can be obtained easily for an arbitrary dimensional state space. For higher moment truncations there is normally a practical upper limit[5] on dimension

[5]This restriction is caused by the need to perform accurate multidimensional integrals of the

associated with the necessity of performing the optimization of distributions.

The mathematical machinery developed was then applied to one of the simplest models of mid-latitude large scale turbulence namely a two-level quasi-geostrophic model with constant vertical shear on a beta plane. Such a configuration simulates reasonably well the generation of turbulence through baroclinic instability and has a cascade from these baroclinic perturbations to dominant barotropic large scale modes which roughly approximates that thought to occur in the real atmosphere and ocean. The equilibrium configuration for this cascade shows typically that most energy is concentrated in the large scale barotropic modes. This effect tends to be significantly greater in flows with small rather than large Rossby radius. Nevertheless the first four Fourier modes typically explained a considerable fraction of the point-wise variation of the stream function at both vertical levels. Motivated by this we simplified the predictability analysis by confining our attention in this study to this highly reduced state space. We plan to extend the dimension of the reduced space in further studies and also consider quantities such as the mid-level temperature field which depends on the (neglected) baroclinic part of the flow.

A question of some practical importance in weather and ocean prediction concerns how predictability varies from one forecast to another and what is the dominant control on such variations. Often one views skill spread diagrams under the assumption that variations in ensemble spread are the dominant control over predictability. The machinery we have developed here allows us to address these issues from a somewhat more fundamental viewpoint.

When a representative sample of initial conditions are chosen and the utility calculated using the geometric partitioning strategy there are often quite large variations at most prediction times (Figure 3) and these variations are often not particularly well related to ensemble spread changes. On the other hand they can be strongly related to a quantity referred to in previous publications by the authors as the signal. This is derived from the Gaussian expression for relative entropy and involves the difference in first moments of the prediction and climatology scaled by the climatological covariance matrix.

At longer prediction times it is often the case that the prediction distributions exhibit significant non-Gaussian behavior which is associated with the relaxation of the reduced state space $p.d.f.s$ around a energy conserving sphere. A natural question to ask is how important are such effects to variations in predictability. We used the moment truncation strategy to assess this but for computational reasons restricted ourselves to examining only the univariate case where the optimization process for $p.d.f.$ computation is very straightforward and well understood in the literature. Our conclusion based on this restricted study was that the variations in relative entropy associated with first two moments tended to dominate that associated with the third and fourth moments despite the evident non-Gaussianicity of the univariate distributions. We intend repeating this calculation in the near future with a multivariate estimator for the fourth

distributions and also their moments. It might be expected that up to about ten dimensions may be feasible.

moment truncated entropy.

Appendix: Moment truncation estimation methods

We describe firstly the estimation of *p.d.f.s* consistent with a truncated set of moments. For this problem we follow the well known maximum entropy principle approach for *p.d.f.* estimation (see e.g., BS p207, Mead and Papanicolaou [20] and MKC). Here one determines the *p.d.f.* with "minimum information content" which is consistent with the constraints that it has moments up to a given truncation point which are the same as those determined from the ensemble. It is very well known that "maximal ignorance" with respect to the higher order neglected moments implies that the *p.d.f.* must belong to the exponential family of distributions. Using standard techniques from the calculus of variations it is straightforward to show that that the *p.d.f.* p under consideration has the form

$$\ln\left(p(\mathbf{x},\overrightarrow{\theta})\right) = -\sum_{i=1}^{L}\{\theta_i h_i(\mathbf{x})\} - \ln Z(\overrightarrow{\theta}) \tag{9}$$

where the $h_i(\mathbf{x})$ are polynomials corresponding to the retained L moments[6] while the parameters $\overrightarrow{\theta}$ are chosen so that the moment constraints are satisfied. In other-words so that

$$M_i \equiv \overline{h_i(\mathbf{x})} = \int h_i(\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{10}$$

where the over-bar indicates a sample mean of the given polynomial. The function $Z(\overrightarrow{\theta})$ is a normalization function of the *p.d.f.* parameters sometimes referred to as the partition function. It is convenient to define an associated function

$$\Gamma(\overrightarrow{\theta}) \equiv \ln Z(\overrightarrow{\theta}) + \sum_{i=1}^{L}\theta_i M_i \tag{11}$$

which is often referred to as the Legendre potential because the relation between the $M_i$and the θ_i is given by the so-called Legendre transformation (see Amari and Nagaoka [2] for a good overview and insight into the connection with differential geometry). It is easily seen that the Legendre potential is minimized by the parameter set $\overrightarrow{\theta} = \overrightarrow{G}$ which ensures that the moment constraints of equation (10) are satisfied. Further one may also easily see that for a general member of the exponential family described here that

$$D(p(\mathbf{x},\overrightarrow{G}),p(\mathbf{x},\overrightarrow{\theta})) = \Gamma(\overrightarrow{\theta}) - \Gamma(\overrightarrow{G}) \tag{12}$$

and thus that the Legendre potential function describes the relative entropy functional on the exponential family. It is clear that once $\overrightarrow{G}$ is computed that

[6] Thus for example $h_i(\mathbf{x}) = x^2$ corresponds with a retained moment of $\left\langle x^2\right\rangle$ and so on.

the utility can then be computed using equation (11) assuming that we retain the same moment set for the climatological and prediction *p.d.f.s*.

The expected information loss due to sampling may also be calculated using the above machinery when combined with some standard results from Bayesian analysis (see BS, Chapter 5 for a good overview of this latter subject). The distribution given in equation (9) is referred to in Bayesian terminology as the likelihood function and one deduces the so-called posterior function using Bayes theorem in the same way as was done in section 2.1 for the *p.d.f.* $P(\mathbf{p}|\mathbf{f})$. The posterior function is the probability that a particular member of the exponential family $p(\mathbf{x}, \overrightarrow{\theta})$ will represent the true *p.d.f.*assuming that such a *p.d.f.*must come from such a family and that the sample moments are M_i. If one assumes that ensemble members are i.i.d[7] according to $p(\mathbf{x}, \overrightarrow{\theta})$ then one may establish using Bayes theorem and an assumption of a uniform prior that the probability of the parameter set $\overrightarrow{\theta}$ given the ensemble moments $M_i$is

$$Q(\overrightarrow{\theta}|\overrightarrow{M}) = \frac{\exp\left[-n\sum_{i=1}^{L}\theta_i M_i\right]}{Z^n(\overrightarrow{\theta})R(\overrightarrow{M})}$$

where

$$R(\overrightarrow{M}) = \int d\overrightarrow{\theta}\exp\left[-n\sum_{i=1}^{L}\theta_i M_i\right]$$

and where the ensemble size is n. The expected information loss then in assuming a particular $p(\mathbf{x}, \overrightarrow{\theta})$ given the observed M_i is thus

$$EL(\overrightarrow{\theta}) = \int d\overrightarrow{\lambda}Q(\overrightarrow{\lambda}|\overrightarrow{M})D(p(\mathbf{x}, \overrightarrow{\lambda}), p(\mathbf{x}, \overrightarrow{\theta}))$$

Straightforward calculation now gives

$$EL(\overrightarrow{\theta}) = \ln Z(\overrightarrow{\theta}) + \sum_{i=1}^{L}\theta_i\left\langle m_i(\overrightarrow{\lambda})\right\rangle_Q - \left\langle \ln Z(\overrightarrow{\lambda}) + \sum_{i=1}^{L}\lambda_i m_i(\overrightarrow{\lambda})\right\rangle_Q \tag{13}$$

where

$$m_i(\overrightarrow{\lambda}) \equiv \int h_i(\mathbf{x})p(\mathbf{x}, \overrightarrow{\lambda})d\mathbf{x}$$

are the moments for the *p.d.f.'s* with parameters $\overrightarrow{\lambda}$ and the notation $\langle\rangle_Q$ denotes expectation of a particular function of the parameters $\overrightarrow{\lambda}$ with respect to the posterior distribution Q. Proposition 5.7 from BS shows that the first expectation has the following intuitive value

$$M_i = \left\langle m_i(\overrightarrow{\lambda})\right\rangle_Q$$

[7]Independently and identically distributed

and this implies that the expected information loss will be minimized if we choose $\theta_i = G_i$ where this set of parameters ensures that the moments of the corresponding $p.d.f.$ match those of the sample estimate (q.f. equation 10 above). Evaluation of the second expectation value in equation (13) is more difficult and in the general case not possible analytically. Fortunately simplification and an analytical form is possible in two important cases: When the $p.d.f.'s$ are Gaussian all integrals may be performed (albeit tediously) with the result that the minimum expected information loss is given by

$$EL^{min}_{Gaussian}(m,n) = \frac{m}{2}\left[F_n - \ln\left(\frac{n}{2}\right)\right] \tag{14}$$

where n is the ensemble size and m is the dimension of the space of state variables under consideration. In addition

$$\begin{aligned} F_n &\equiv \quad 1 + \tfrac{1}{2} + \tfrac{1}{3} + ... + \tfrac{2}{n+1} - C \quad n\ odd \\ &\equiv \quad 2(1 + \tfrac{1}{2} + \tfrac{1}{3} + ... + \tfrac{1}{n+1}) - C - \ln 4 \quad n\ even \end{aligned}$$

where $C = 0.577215...$is Euler's constant. As expected

$$\lim_{n\to\infty} EL^{min}_{Gaussian}(m,n) = 0$$

Note also that the parameters for the Gaussian $p.d.f.'s$ which minimize the expected information loss due to sampling are precisely those expected intuitively i.e.

$$p(\mathbf{x}) \propto \exp\left(-(\mathbf{x} - \overrightarrow{\mu})^t \Sigma^{-1} (\mathbf{x} - \overrightarrow{\mu})\right) \tag{15}$$

where $\overrightarrow{\mu}$ is the sample mean vector and Σ is the sample covariance matrix.

The other situation where analytical results are possible is the asymptotic case i.e. when the ensemble is sufficiently large so that the central limit theorem may be used to reduce the posterior to a normal distribution which is tightly centered on the EL minimization set of parameters. This scenario has been extensively (and rigorously) studied in the statistics literature and for the exponential family of $p.d.f.'s$ one may show (BS, Proposition 5.16) that asymptotically the posterior function is normal with mean vector G_i and covariance matrix

$$\Sigma_n = \frac{1}{n} H^{-1}(\overrightarrow{G}) \tag{16}$$

where the matrix $H(\overrightarrow{G})$ is called the Hessian matrix (evaluated at $\overrightarrow{G}$) and is given by

$$H_{ij}(\overrightarrow{G}) = \left.\frac{\partial\left(\ln Z(\overrightarrow{\lambda})\right)}{\partial\lambda_i \partial\lambda_j}\right|_{\overrightarrow{\lambda}=\overrightarrow{G}} \tag{17}$$

The form of the posterior function implies that only $p.d.f.s$ that are "relatively close" in parameter space make a contribution to the expected information loss. This follows from the $\frac{1}{n}$ dependence of the covariance function Σ_n which reflects the fact that we are more certain of the parameters $\overrightarrow{\lambda}$ the larger the sample size n. This allows us to make a very useful approximation for the relative entropy which is much discussed in literature on information geometry (see e.g. Amari and Nagaoka [2] pp54-55):

$$D(p(\mathbf{x}, \overrightarrow{\lambda}), p(\mathbf{x}, \overrightarrow{\lambda} + \delta\overrightarrow{\lambda})) \simeq H_{ij}(\overrightarrow{\lambda})\delta\lambda_i\delta\lambda_j$$

where the Hessian matrix H_{ij} is now playing the role of a Riemannian metric tensor and we are assuming the summation convention. Note that when the $p.d.f.s$ are "close" in parameter space this approximation implies that the relative entropy acts like the usual distance function. If we insert this approximation into equation (13) and use the Gaussian form of the posterior distribution discussed above (see equation (16)), we obtain the following very simple expression relevant to large samples:

$$EL = \frac{m}{n} \tag{18}$$

in other words the larger the sample the lower the loss and the greater the number of retained moments the greater the loss.

A strategy for calculating the low order moment information content of a prediction ensemble and the expected loss due to sampling can now be outlined: If only the first and second moment information is used then one uses equations (14) and (15) combined with the analytical form for relative entropy of Gaussian $p.d.f.'s$. If one has a sufficiently large ensemble to hand so that the information content of third and higher moments might be reliably used then one minimizes the Legendre potential (equation 11) with respect to the parameter set $\overrightarrow{\theta}$ using an unconstrained non-linear optimization technique. Once $p.d.f.$ estimates for the climatological and prediction distribution are obtained, equations (12) and (18) can be used to calculate the utility and the expected sampling information loss.

References

[1] R. Abramov and A. J. Majda, *Quantifying uncertainty for non-gaussian ensembles in complex systems*, SIAM J. Sci. Stat. Comput. (2003), In press.

[2] S. Amari and H. Nagaoka, *Methods of information geometry*, Translations of Mathematical Monographs, AMS, Oxford University Press, 2000.

[3] J. Bernardo and A. Smith, *Bayesian theory*, John Wiley and Sons, 1994.

[4] L. Boltzmann, *Lectures on gas theory*, Dover, March 1995.

[5] R. Buizza and T. N. Palmer, *Impact of ensemble size on ensemble prediction*, Mon. Weath. Rev. **126** (1998), 2503–2518.

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

[7] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley, New York, New York, 1991.

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

[9] E. S. Epstein, *The role of initial uncertainties in prediction*, J. Appl. Met. **8** (1969), 190–198.

[10] J. S. Frederiksen, *Nonlinear stability of baroclinic flows over topography*, Geophys. Astrophys. Fluid Dynamics **57** (1991), 85–97.

[11] P. Grassberger, *Generalized dimensions of strange attractors*, Phys. Lett. A **97** (1983), 227–230.

[12] I. M. Held and V. D. Larichev, *A scaling theory for horizontally homogeneous, baroclinically unstable flow on a beta plane*, J. Atmos. Sci. **53** (1996), 946–952.

[13] P. L. Houtekamer, L.Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, *A system simulation approach to ensemble prediction*, Mon. Weath. Rev. **124** (1996), 1225–1242.

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

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

[16] C. E. Leith, *Theoretical skill of monte carlo forecasts*, Mon. Weath. Rev. **102** (1974), 409–418.

[17] E. N. Lorenz, *Deterministic non-periodic flows*, J. Atmos. Sci. **20** (1963), 130–141.

[18] A. J. Majda, R. Kleeman, and D. Cai, *A framework for predictability through relative entropy*, Meth. Appl. Anal. (2003), In press.

[19] A. J. Majda and I. Timofeyev, *Remarkable statistical behavior for truncated burgers-hopf dynamics*, Proc. Nat. Acad. Sci. USA **97** (2000), 12413–12417.

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

[21] F. Molteni, R. Buizza, T. N. Palmer, and T. Petroliagis, *The ecmwf ensemble prediction system: Methodology and validation*, Quart. J. Roy. Meteor. Soc. **122** (1996), 73–119.

[22] J. M. Murphy, *The impact of ensemble prediction on predictability*, Quart. J. Roy. Met. Soc. **114** (1988), 299–323.

[23] T. N. Palmer, F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, *Ensemble prediction*, Proc. Validation of Models over Europe **1** (1993), 21–66.

[24] T. N. Palmer and S. Tibaldi, *On the prediction of forecast skill*, Mon. Weath. Rev. **116** (1988), 2453–2480.

[25] M. S. Roulston and L. A. Smith, *Evaluating probabilistic forecasts using information theory*, Mon. Weath. Rev. **130** (2002), 1653–1660.

[26] D. Ruelle and F. Takens, *On the nature of turbulence*, Comm. Math. Phys. **20** (1971), 167–192.

[27] R. Salmon, *Two-layer quasi-geostrophic turbulence in a simple special case*, Geophys. Astrophys. Fluid Dyn. **10** (1978), 25–52.

[28] ———, *Baroclinic instability and geostrophic turbulence*, Geophys. Astrophys. Fluid Dyn. **15** (1980), 167–211.

[29] ———, *Lectures on geophysical fluid dynamics*, Oxford Univ. Press, New York, 1998.

[30] T. Schneider and S. Griffies, *A conceptual framework for predictability studies*, J. Clim. **12** (1999), 3133–3155.

[31] S. K. Smith, G. Boccaletti, C. C. Henning, I. N. Marinov, C. Y. Tam, I. M. Held, and G. K. Vallis, *Turbulent diffusion in the geostrophic inverse cascade*, J. Fluid Mech. **469** (2002), 13–48.

[32] Z. Toth and E. Kalnay, *Operational ensemble prediction at the national meteorological center. practical aspects*, Bull. Am. Meteorol. Soc. **74** (1993), 2317–2330.

[33] ———, *Ensemble forecasting at ncep and the breeding method*, Mon. Weath. Rev. **125** (1997), 3297–3319.