A Mathematical Framework for Quantifying Predictability Through Relative Entropy

by
Andrew Majda
Courant Institute of Mathematical Sciences
Center for Atmosphere Ocean Science
New York University
Richard Kleeman
Courant Institute of Mathematical Sciences
Center for Atmosphere Ocean Science
New York University
and
David Cai
Courant Institute of Mathematical Sciences

Abstract

Kleeman has recently demonstrated that the relative entropy provides a significant measure of the information content of a prediction ensemble compared with the climate record in several simplified climate models. Here several additional aspects of utilizing the relative entropy for predictability theory are developed with full mathematical rigor in a systematic fashion which the authors believe will be very useful in practical problems with many degrees of freedom in atmosphere/ocean and biological science. The results developed here include a generalized signal-dispersion decomposition, rigorous explicit lower bound estimators for information content, and rigorous lower bound estimates on relative entropy for many variables, N, through N, one-dimensional relative entropies and N, two-dimensional mutual information functions. These last results provide a practical context for rapid evaluation of the predictive information content in a large number of variables.

1 Introduction

Donsker and Varadhan (see the research-expository article by Varadhan 1984, 1985 for many additional references) have made profound use of the entropy and relative entropy in their studies of large deviations in probability theory. Following the pioneering work by H.T. Yau (1991), Olla, Varadhan, and Yau (1993) have utilized relative entropy methods to prove interesting theorems on hydrodynamic limits of various particle systems. In this paper, we develop a mathematical framework for predictability utilizing the relative entropy. While no difficult rigorous mathematical results are presented here, the authors believe the systematic framework developed below will be very useful in quantifying the predictability and information content in ensemble predictions for highly inhomogeneous dynamical systems arising in diverse practical applications such as climate or weather prediction and biological molecular dynamics. In the remainder of the introduction, we present a brief discussion to these topics as well as the role of relative entropy in quantifying predictability and also briefly outline the remainder of the paper.

1.1 Ensemble Prediction and Relative Entropy as a Measure of Predictability

For chaotic dynamical systems as usually occur in modeling geophysical flows or other complex applications, one typically is not only interested in the behavior of an individual solution but the statistical behavior of an entire ensemble of solutions starting from nearly the same initial data. We illustrate this behavior next in an idealized setting. To make matters concrete consider a large system of stochastic ODE's for a vector $\vec{X} \in R^N$, $N \gg 1$, given by

$$d\vec{X} = \vec{F}(\vec{X}(t), t)dt + \vec{g}(\vec{X}(t), t)d\vec{W} \tag{1.1}$$

where $\vec{F}$ is a vector-field of dimension N, $\vec{W}$ is a standard K-dimensional Wiener process, and g is an $N \times K$ matrix valued function. Examples with the structural form in (1.1) are abundant both for geophysical flows (Majda and Wang, 2002; Kleeman, 2000) and for biomolecular dynamics (Schutte, Fisher, Huisinga, Deuflhard, 1999).

Consider a deterministic initial data for (1.1), $\vec{X}_0$, and an associated probability density consisting of small random perturbations of this initial data. This probability density measures the uncertainty in the measurement of the initial data. For example, the random initial data might be sampled from a Gaussian probability distribution centered about $\vec{X}_0$,

$$p_0(\vec{X}) = (2\pi)^{-N/2}\epsilon^{-N/2}\, e^{-\frac{|\vec{X}-\vec{X}_0|^2}{2\epsilon}}, \epsilon \ll 1 \tag{1.2}$$

where $\epsilon \ll 1$ measures the variance in each component. In other words, the ensemble of initial data satisfies

$$\begin{aligned}
&\int \vec{X} p_0(\vec{X}) d\vec{X} = \vec{X}_0, \\
&\int (X_i - X_{0,i})(X_j - X_{j,0}) p_0(\vec{X}) d\vec{X} = \epsilon \delta_{ij}
\end{aligned} \tag{1.3}$$

where δ_{ij} are the Kronecker symbols and take the value 1 if $i = j$ and zero otherwise. In general, here we assume an initial probability measure $p_0(\vec{X})$, reflecting an ensemble of initial conditions. The initial probability density evolves to a new probability density $p_t(\vec{X})$ for $t \geq 0$ which satisfies the Fokker–Planck equation

$$\begin{aligned}
&\frac{\partial}{\partial t} p_t(\vec{X}) = -\mathrm{div}_{\vec{X}}(\vec{F} p_t) + \sum_{i,j} \frac{\partial^2((Q/2)_{ij} p_t)}{\partial X_i \partial X_j} \\
&p_t(\vec{X})|_{t=0} = p_0(\vec{X})
\end{aligned} \tag{1.4}$$

where $Q(\vec{X}, t)$ is the $N \times N$ matrix, $Q = g(\vec{X}, t) g^T(\vec{X}, t)$. Of course, in practice, it is impractical to solve the Fokker–Planck equation in (1.4) for $N \gg 1$ and instead a large ensemble of solutions is generated by sampling the initial distribution, $p_0(\vec{X})$, and generating individual solutions by solving (1.1) for each initial data and then building an approximation to the ensemble mean and typically the first few moments of $p_t(\vec{X})$, i.e.

$$\begin{aligned}
&\int X_i p_t(\vec{X}) d\vec{X} = \vec{X}_{t,i}, \\
&\int (X_i - \overline{X}_{t,i})^\alpha (X_j - \overline{X}_{t,j})^\beta \, p_t(\vec{X}) d\vec{X} = M_{i,j}^{\alpha,\beta}(t)
\end{aligned} \tag{1.5}$$

for $|\alpha| + |\beta| \leq 2L$. Such an approach is feasible with the current generation of supercomputers provided that the time of integration is not unrealistically long. For example, in current weather prediction models for mid-latitudes, on the order of 50 different realizations are utilized while intermediate models for predicting El Nino involve ensemble prediction with as many as 500 different realizations (Kleeman (2000)). In practice, only the low order moments for a suitable collection of variables can be measured with statistical significance from such an ensemble, i.e.,

$$\begin{aligned}
&X_{t,i} \text{ is measured for } 1 \leq i \leq M, \\
&M_{i,j}^{\alpha,\beta} \text{ is measured for } 1 \leq i, j \leq M \text{ and } |\alpha| + |\beta| \leq 2L \text{ with } L = 2 \text{ or } L = 4.
\end{aligned} \tag{1.6}$$

1.1.1 Relative Entropy as a Measure of Predictive Information Content

How much information does an ensemble prediction have beyond the historical climate record? How can this be quantified? Consider a subset X_i, $1 \leq i \leq M$ of the variable $\vec{X}$ defining the full dynamics in (1.1). For this collection of variables X_i, the historical climate record is a probability measure, $\Pi(X_1, \ldots, X_M)$, which can be regarded as known. When will a short term ensemble prediction be useful and contain more information beyond the climate record? In an important paper, Kleeman (2000) has suggested that given a probability measure $p(X_1, \ldots, X_M)$ for an ensemble prediction at a given time, the relative entropy quantifies this additional information. In other words, the number $P(p, \Pi)$ where

$$P(p,\Pi) = -\mathcal{S}(p,\Pi) = H(p,\Pi) = \int_{\mathcal{R}^M} p \ln\left(\frac{p}{\Pi}\right) \tag{1.7}$$

precisely measures the information content (Cover and Thomas, 1991) in the prediction ensemble p, beyond that in the historical climate record determined by Π. Important mathematical support for the practical significance of relative entropy is given by the property,

$$P(p,\Pi) > 0 \text{ unless } p = \Pi. \tag{1.8}$$

Furthermore, for prediction ensembles which satisfy the stochastic equations in (1.1), under suitable additional hypotheses it is known (Cover and Thomas (1991), Lasota and Mackey (1994)) that

$$\begin{aligned} &P(p_t,\Pi) \text{ decreases in time, and} \\ &P(p_t,\Pi) \to 0, \text{ as } t \to \infty. \end{aligned} \tag{1.9}$$

For a general probability density, $p(X_1, \ldots, X_M)$, the mean and covariance matrix, Cov$p$, are defined by

$$\begin{aligned} \overline{X}_i &= \int X_i p, \quad 1 \le i \le M, \\ (\mathrm{Cov}p)_{ij} &= \int (X_i - \overline{X}_i)(X_j - \overline{X}_j)p. \end{aligned} \tag{1.10}$$

Note that Cov$p$ is a symmetric positive definite matrix since for any vector $\xi \in \mathcal{R}^M$,

$$\sum (\mathrm{Cov}p)_{ij}\xi_i\xi_j = \int \left(\sum_{i=1}^{M} \xi_i (X_i - \overline{X}_i)\right)^2 p > 0 \tag{1.11}$$

provided that $\xi \neq 0$ and p is not a delta function.

1.1.2 Practical and Mathematical Issues for Predictability

Some important issues regarding predictability are the following:

(1.12)

A) How much do the respective information in the mean and/or the covariance of both the prediction, p, and climate distribution, Π, influence the information content in a prediction, $P(p, \Pi)$, for a given set of variables in a given dynamical system? How can this be quantified?

B) In a given dynamical system, which subset of variables, $X_1, \ldots, X_M$ are more predictable than others? How can this be quantified?

C) How can one practically compute or approximate the relative entropy, $P(p, \Pi)$ for a complex many degree of freedom system and a subset of variables with $M \gg 1$?

D) Are there mathematical strategies to rigorously estimate, $P(p, \Pi) \geq P(p_*, \Pi)$, in a given context where $P(p_*, \Pi)$ can be evaluated either analytically or by a rapid numerical procedure?

All of these important practical issues will be addressed below through elementary mathematical ideas.

1.1.3 Gaussian Climates

In many practical applications (Toth, 1991; Schneider and Griffies 1999, Kleeman, 2000), the overall dynamics associated with (1.1) is extremely complex and chaotic but nevertheless the climate distribution is essentially Gaussian on a suitable subset of variables and thus is completely determined by its mean and covariance. For such a Gaussian distribution,

$$\Pi_0(X_1, \ldots, X_M) = (2\pi)^{-M/2} (\det \mathcal{C})^{-1/2} \exp\left(-\frac{1}{2}((\vec{X} - \vec{X}_0), \mathcal{C}^{-1}(\vec{X} - \vec{X}_0)) \right) \tag{1.13}$$

where $\vec{X}_0 \in \mathcal{R}^M$ is the mean,

$$\int X_j \Pi_0 = X_{0,j}, \qquad 1 \leq j \leq M \tag{1.14}$$

and $\mathcal{C} = (C_{ij})$ is the positive definite, $\mathcal{C} > 0$, symmetric, $M \times M$ covariance matrix,

$$\int (X_i - X_{0,i})(X_j - X_{0,j})\Pi_0 = C_{ij}. \tag{1.15}$$

The practical and mathematical issues for predictability listed in (1.12) are most readily developed for the special situation where $\Pi(X_1, \ldots, X_M)$ is defined by a Gaussian climate. The theory for this important special case is developed in Section 2 below. The more general theory for the practical and theoretical issues in (1.12) is developed in Sections 3 and 4 with the case of a Gaussian climate as a pedagogical guideline. The important practical issue in (1.12) C) is addressed in Section 4. Furthermore, many systems of interest as important models for geophysical flows have explicit Gaussian climate distributions, (Majda, Timofeyev, Vanden Eijinden, 2001; Majda and Timofeyev, 2000; Majda and Wang, 20002).

1.1.4 Invariance of the Predictability Measure Under a General Change of Coordinates

Clearly, a good predictability measure for features of a given dynamical system should not depend on the coordinate system used to describe the underlying dynamics. While this seems obvious intuitively, many ad-hoc predictability measures utilized in practice fail to have this property and often depend on the choice of metric (energy, geopotential height etc.); Schneider and Griffies (1999) have emphasized this point in their work. Here we show that the predictability measure, $P(p, \Pi)$, defined in (1.7) by the relative entropy is invariant under a general nonlinear change of coordinates.

Consider a smooth invertible transformation from $\mathcal{R}^M$ to $\mathcal{R}^M$ defined by

$$\vec{Y} \to \Phi(\vec{Y}) = \vec{X}.$$

Any smooth probability density, $p(\vec{X})$, determines a smooth probability density $p_\Phi(\vec{Y})$ via the formula

$$p_\Phi(\vec{Y}) = p(\Phi(\vec{Y})) \left| \det \left(\frac{d\Phi}{d\vec{Y}} \right) \right| \tag{1.16}$$

where $\frac{d\Phi}{d\vec{Y}}$ denotes shorthand for the $M \times M$ Jacobian matrix of the mapping, $\frac{d\Phi}{d\vec{Y}} = \left(\frac{\partial \Phi_i}{\partial Y_j} \right)$ and det is the determinant.

The standard change of variable formula shows that if $p(\vec{X})$ is a probability density then $p_\Phi(\vec{Y})$ is also a probability density: the factor det $\left(\frac{d\Phi}{d\vec{Y}} \right)$ is needed to guarantee this. The invariance of the relative entropy predictability measure under change of coordinates is the statement that

$$\int p \ln \left(\frac{p}{\Pi} \right) = P(p, \Pi) = P(p_\Phi, \Pi_\Phi) = \int p_\Phi \ln \left(\frac{p_\Phi}{\Pi_\Phi} \right) \tag{1.17}$$

for any smooth invertible transformation $\Phi(\vec{Y})$.

However, it follows immediately from (1.16) and the change of variables formula that (1.17) is true.

2 Quantifying Predictability for Gaussian Climate Variables

First as a simple application of the principle in (1.17) we show how measuring predictability for a Gaussian climate can be simplified through the special use of a linear change of coordinates. For the Gaussian climate defined in (1.13), consider the change of variables

$$\vec{X} = \vec{X}_0 + \mathcal{C}^{1/2} \vec{Y}. \tag{2.1}$$

In (2.1), $\mathcal{C}^{1/2}$, is the square root of the positive definite matrix $\mathcal{C}$; the matrix $\mathcal{C}^{1/2}$ is also symmetric and positive definite, commutes with $\mathcal{C}$, and satisfies $\mathcal{C}^{1/2}\mathcal{C}^{-1}\mathcal{C}^{1/2} = I$. Since $\mathcal{C}$ is positive definite symmetric, there is an orthonormal basis, $\{\vec{e}_{i=1}^{M}\}$, called empirical orthogonal functions (EOF's) in the atmosphere/ocean community and corresponding positive eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M > 0$ so that the covariance matrix $\mathcal{C}$ is diagonalized, i.e.

$$\mathcal{C}\vec{f} = \sum_i \lambda_i \vec{e}_i(\vec{f}, \vec{e}_i). \tag{2.2}$$

The eigenvectors, $\vec{e}_1$, is called the first EOF, etc. With (2.2), the matrix $\mathcal{C}^{1/2}$ is defined by

$$\mathcal{C}^{1/2}\vec{f} = \sum_i \lambda_i^{1/2} \vec{e}_i(\vec{f}, \vec{e}_i). \tag{2.3}$$

From (1.16) and (1.13) it follows that in the coordinate system defined by (2.1), the Gaussian climate distribution has the simplified form

$$\Pi_0(\vec{Y}) = (2\pi)^{-M/2} e^{-\frac{1}{2}|\vec{Y}|^2}. \tag{2.4}$$

With (1.16) and (2.1), given $p(\vec{X})$, one gets

$$p_\Phi(\vec{Y}) = p(\vec{X}_0 + \mathcal{C}^{1/2}\vec{Y}) \det(\mathcal{C}^{1/2}). \tag{2.5}$$

Given the probability density $p(\vec{X})$ with mean $\vec{X}_0$ and covariance matrix, Cov$p$, defined in (1.10), it follows easily from (2.1), (2.3) that the mean and covariance of the transformed probability density, p_Φ, are given by

$$\begin{aligned} \text{mean } p_\Phi &= \overline{\vec{Y}} = \mathcal{C}^{-1/2}(\overline{\vec{X}} - \overline{\vec{X}}_0), \\ \text{Cov}p_\Phi &= \mathcal{C}^{-1/2} \ \text{Cov}p \ \mathcal{C}^{-1/2}. \end{aligned} \tag{2.6}$$

Next we address the practical and mathematical issues elucidated in 1.1.2 regarding predictive information content for the special case of a Gaussian climate distribution. Here as shown above in (2.1)–(2.5) without loss of generality we assume $\vec{X} = (X_1, \ldots, X_M)$ and

$$\Pi_0(\vec{X}) = (2\pi)^{-M/2} e^{-\frac{1}{2}|\vec{X}|^2}. \tag{2.7}$$

First we calculate the relative entropy of $p(\vec{X})$ which measures the additional information in p beyond Π_0. For simplicity in notation, here we suppress the preliminary transformation Φ that leads to the canonical form in (2.7). Recall that the entropy of a probability density, p, is given by

$$\mathcal{S}(p) = -\int_{\mathcal{R}^M} p \ln p \tag{2.8}$$

which measures the average lack of information in p (Cover and Thomas, 1991; Majda and Wang, 2002).

With (2.7), we calculate

$$P(p,\Pi_0) = \int p\ln p - \int p\ln\Pi_0$$
$$= -\mathcal{S}(p) + \frac{M}{2}\ln(2\pi) + \int_{\mathcal{R}^M}\frac{|\vec{X}|^2}{2}p \tag{2.9}$$

while for the Gaussian distribution in (2.7) we have the explicit formula

$$\mathcal{S}(\Pi_0) = \frac{M}{2}\ln(2\pi) + \frac{M}{2}. \tag{2.10}$$

Thus, we obtain the *Relative entropy identity for a normalized Gaussian climate in* $\mathcal{R}^M$,

$$P(p,\Pi_0) = \mathcal{S}(\Pi_0) - \mathcal{S}(p) + \int_{\mathcal{R}^M}\frac{|\vec{X}|^2}{2}p - \frac{M}{2}. \tag{2.11}$$

The elementary identity in (2.11) states that except for an interesting correction involving second moments, the measure of predictability through relative entropy for a Gaussian climate is directly related to the entropy difference. In fact, an immediate corollary of (2.11) is the following:

Assume the ensemble prediction probability density p satisfies
$$\int |\vec{X}|^2 p = M = \int |\vec{X}|^2\Pi_0, \text{ then } P(p,\Pi_0) = \mathcal{S}(\Pi_0) - \mathcal{S}(p). \tag{2.12}$$

Thus, if the trace of the second moments of the probability density p coincides with that of the climate, the relative entropy is exactly the entropy difference.

2.1 The Signal and Dispersion Decomposition for a Gaussian Climate

For the Gaussian climate in (2.7), here we develop a simple decomposition of the predictability measure $P(p,\Pi_0)$ into signal and dispersion components. This provides a theoretical framework for addressing the issues in (1.12) A). Roughly speaking, the signal measures the contribution of the mean to the information content of the prediction beyond the climate while the dispersion measures the information content of the variance and other high moments for the prediction. Given the fact that the climate is normalized to have zero mean, the knowledge that the prediction ensemble has a non-zero mean has potentially significant information content beyond the normalized variance of the prediction. Measures of predictability involving only the variance alone were the usual concept in atmosphere/ocean science (Schneider and Griffies (1999)) until the important paper of Kleeman (2000) who introduced the signal-dispersion decomposition for the special case when both $p(\vec{X})$ and

$\Pi_0(\vec{X})$ are Gaussian distributions. In that paper, Kleeman (2000) also demonstrated the practical utility of the signal component in predictions for a series of stochastic and deterministic models for El Nino while Kleeman, Majda and Timofeyev (2002) demonstrated the surprising importance of the signal component in determining predictive utility for the truncated Burgers–Hopf models which are one-dimensional models with Gaussian climates and statistical features resembling those for the atmosphere (Majda and Timofeyev, 2000; Majda and Wang, 2002).

For a Gaussian climate, the signal-dispersion decomposition is an immediate consequence of the formula in (2.11) and the elementary identity

$$\int |\vec{X}|^2 p = \sum_{i=1}^{M} \left(\int |\overline{X}_i|^2 p + \int (X_i - \overline{X}_i)^2 p \right) = |\text{mean } p|^2 + tr(\text{Cov}p) \tag{2.13}$$

where given an $M \times M$ matrix $A = (A_{ij})$, the trace of $A, tr(A) = \sum\limits_{1 \leq i \leq M} A_{ii}$. With (2.11) and (2.13), we have the following

Proposition 2.1. (The Signal Dispersion Decomposition for Gaussian Climate Variables) *Under these circumstances $P(p, \Pi_0)$ admits the signal-dispersion decomposition,*

$$P(p, \Pi_0) = S + D \tag{2.14}$$

where the signal S is given by

$$S = \frac{1}{2}|mean\ p|^2 \tag{2.15}$$

and the dispersion, D, is given by

$$D(p, \Pi_0) = \mathcal{S}(\Pi_0) - \mathcal{S}(p) + \frac{1}{2}(tr(\text{Cov}p) - tr(\text{Cov}\Pi_0)). \tag{2.16}$$

The formulas in (2.14), (2.19) and (2.16) are an immediate consequence of (2.11), (2.13) and the identity, $tr(\text{cov } \Pi_0) = M$. Clearly, only a non-zero mean value of p contributes to the information content of the prediction as measured by the signal. On the other hand, only higher moments of p contribute to the information content in the dispersion, D, as defined in (2.16). Note that both $\mathcal{S}(p)$ and $tr(\text{Cov}p)$ have identical values for p and any translated distribution with an arbitrary mean, $p_{\vec{\tau}} = p(\vec{X} - \vec{\tau})$ for a constant vector $\vec{\tau}$; thus, for the dispersion, D

$$D(p_{\vec{\tau}}, \Pi_0) = D(p, \Pi_0) \text{ for any } \vec{\tau} \tag{2.17}$$

and the dispersion in (2.16) measures the information content in p beyond Π_0 which is independent of the value of the mean of p. These facts provide the intuition behind the signal-dispersion decomposition in Proposition 2.1. This decomposition is generalized to suitable non-Gaussian climate distributions in Sections 3 and 4 below.

2.2 Rigorous Lower Bounds on Predictive Information Content with a Gaussian Climate

As noted in (1.5), (1.6), in practical ensemble prediction the mean and a few moments up to order typically at most four of the predictive probability density, $p(\vec{X})$, are known with significant accuracy. Here we provide a rigorous mathematical framework for estimating the predictive information content by quantities which can be readily computed. Thus, we address the issues in C) and D) from (1.12) for the special case of a Gaussian climate as in (2.7). A significant straightforward generalization for a non-Gaussian climate is presented below in Section 3.

Motivated by (1.5), (1.6), we assume that we know the mean and a finite number of the higher moments for $p(\vec{X})$, i.e.

$$\begin{aligned}\overline{X}_i &= \int X_i p, \qquad 1 \le i \le M, \\ M_{i,j}^{\alpha,\beta} &= \int (X_i - \overline{X}_i)^\alpha (X_j - \overline{X}_j)^\beta p, \qquad 0 \le \alpha+\beta \le 2L, 1 \le i,j \le M, \text{ with } L \ge 1 \text{ fixed.}\end{aligned} \tag{2.18}$$

The issues we address here are how can we estimate and rigorously bound $P(p,\Pi_0)$ in a systematic and practically significant fashion?

As background information, recall that the Gaussian distribution maximizes the entropy among all probability measures with specified first and second moments. Recall from (2.14)–(2.16) that

$$P(p,\Pi_0) = -\mathcal{S}(p) + R(p,\Pi_0) \tag{2.19}$$

where R depends solely on the first and second moments of p, i.e.

$$R(p,\Pi_0) = \mathcal{S}(\Pi_0) + \frac{1}{2}(tr(\text{Cov}p) - tr(\text{Cov}\Pi_0)) + \frac{1}{2}(\text{mean } p)^2. \tag{2.20}$$

From (2.19) it follows that if we pick a probability density, p^*, which maximizes the entropy, $\mathcal{S}(p)$, subject to some of the constraints with $L \ge 1$ so that $R(p,\Pi_0)$ remains constant, then automatically $P(p^*,\Pi_0)$ provides a rigorous and potentially practical lower bound on the information content in the ensemble prediction beyond the climate. Next, we formalize the statements above in a precise fashion.

For any fixed $\tilde{L}$ with $1 \le \tilde{L} \le L$, define a set, $\mathcal{C}_{\tilde{L}}$, of linear constraints on probability measures through all the moment constraints in (2.18) for $\alpha+\beta \le 2\tilde{L}$. By definition, the prediction ensemble, p, belongs to $\mathcal{C}_{\tilde{L}}$ so this set is non-empty and convex while the entropy is a concave function on this set. Thus, we can always find $p^*_{2\tilde{L}}$ which satisfies the *maximum entropy principle*

$$p^*_{2\tilde{L}} \in \mathcal{C}_{\tilde{L}}, \quad \mathcal{S}(p_{2\tilde{L}}) = \max_{p\in\mathcal{L}_{\tilde{L}}} \mathcal{S}(p). \tag{2.21}$$

Clearly, we have

$$\mathcal{S}(p_{2L_1}) \geq \mathcal{S}(p_{2L_2}) \geq \mathcal{S}(p) \quad \text{for any} \quad 1 \leq L_1 \leq L_2 \leq L. \tag{2.22}$$

Furthermore, because $\tilde{L} \geq 1$, and $R(p, \Pi_0)$ only involves moments up to order two,

$$R(p^*_{2\tilde{L}}, \Pi_0) = R(p, \Pi_0) \text{ for any } \tilde{L}, \text{ with } 1 \leq \tilde{L} \leq L \tag{2.23}$$

and $p^*_1 = p^*_G$ is a Gaussian distribution with the same mean and variance matrix as p: thus p^*_G is given by explicit formulas as in (1.13)–(1.15).

(2.24)

With (2.19)–(2.24), we have the following

Proposition 2.2. (Estimating the Information Content in an Ensemble Prediction for Gaussian Climate Variables) *For any fixed number of prediction moments, $2L$, with $L \geq 1$ with a Gaussian climate, we have the rigorous lower bounds*

$$P(p, \Pi_0) \geq P(p^*_{2L}, \Pi_0) \geq P(p^*_{2\tilde{L}}, \Pi_0) \geq P(p^*_G, \Pi_0), \quad \textit{for} \quad L \geq \tilde{L} \geq 1. \tag{2.25}$$

In fact, the same chain of estimates in (2.25) *also applies to the dispersion $D(p, \Pi_0)$, alone, i.e.*

$$D(p, \Pi_0) \geq D(p^*_{2L}, \Pi_0) \geq D(p^*_{2\tilde{L}}, \Pi_0) \geq D(p^*_G, \Pi_0), \quad \textit{for} \quad L \geq \tilde{L} \geq 1. \tag{2.26}$$

The estimates in (2.25) are generalized to a non-Gaussian climate in Section 3 below. Kleeman (2000) introduced and applied the relative entropy, $P(p, \Pi_0)$, for measuring the information content where he assumed that both p and Π_0 were Gaussian distributions. Here in Proposition 2.2, we have shown that the relative entropy of the Gaussian distribution, $P(p^*_G, \Pi_0)$ is automatically a rigorous lower bound on the information content of any prediction density p with the same first and second moments. With the formula in (2.16) for the dispersion and (1.13), for the Gaussian estimator $P(p^*_G, \Pi_0)$, we have

$$\mathcal{S}(\Pi_0) - \mathcal{S}(p^*_G) = \ln(\det(\mathrm{Cov}p)^{-1/2}) \tag{2.27}$$

so that

$$\begin{aligned} D(p^*_G, \Pi_0) &= \ln(\det(\mathrm{Cov}p)^{-1/2}) + \frac{1}{2}(tr(\mathrm{Cov}\ p) - M), \\ P(p^*_G, \Pi_0) &= D(p^*_G, \Pi_0) + \frac{1}{2}|\text{mean } p|^2. \end{aligned} \tag{2.28}$$

Of course $D(p^*_G, \Pi_0)$ is just the relative entropy of the Gaussian measure with the same variance as P^*_G but with zero mean.

Finally we mention that the practical advantage for utilizing the maximum entropy principle in (2.21) with the explicit moment information in (2.18) is that standard constrained optimization numerical methods can be utilized to find $p^*_{2\tilde{L}}$ accurately. Here an even number of moments are utilized to guarantee that the optimization problem has a solution.

2.3 Choosing Reduced Variables to Order the Predictive Information Content

Here we describe a nice construction for a Gaussian prediction distribution, p_G^*, essentially due to Schneider and Griffies (1999) which organizes the variables $\vec{X}$ into subspaces with a hierarchy of predictive information content. Recall that through the change of variables in (2.1), a general Gaussian climate assumes the canonical form in (2.4) while the relative entropy of a general prediction remains invariant; the covariance matrix for a general prediction distribution in the new coordinates is given by the formula in (2.6). For estimating predictive information content, it is very natural to introduce a second change of coordinate which diagonalizes the covariance matrix of p. Since Cov$p$ is a symmetric positive definite $M \times M$ matrix, there exists a rotation matrix, $\mathcal{O}$, with $\mathcal{O}^T = \mathcal{O}^{-1}$, so that

$$\mathcal{O}^{-1}\ \mathrm{Cov}p\ \mathcal{O} = D = \begin{pmatrix} \gamma_1 & 0 & \dots & 0 \\ 0 & \gamma_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \gamma_M \end{pmatrix} > 0 \tag{2.29}$$

where D is the positive diagonal matrix with non-zero diagonal entries, $\gamma_i > 0$. Consider the new variable

$$\vec{X} = \mathcal{O}\vec{Z}. \tag{2.30}$$

Since $\mathcal{O}$ is a rotation matrix, it follows that the Gaussian climate measure $\Pi_0(\vec{Z})$ retains the same normalized form in (2.4).

In this coordinate system, as a consequence of (2.29), the Gaussian distribution $p_G^*(\vec{Z})$ with the same variance and mean as p, assumes the factored form

$$p_G^*(\vec{Z}) = \Pi_{i=1}^M (2\pi\gamma_i)^{-1/2} e^{\frac{(Z_i - \bar{Z}_i)^2}{2\gamma_i}} \stackrel{\text{def}}{\equiv} \Pi_{i=1}^M p_i^*(Z_i) \tag{2.31}$$

while the climate distribution $\Pi_0(\vec{Z})$ retains the factored form

$$\Pi_0(\vec{Z}) = \Pi_{i=1}^M (2\pi)^{-1/2}\, e^{-\frac{Z_i^2}{2}} = \Pi_{i=1}^M \Pi_{0,i}(Z_i). \tag{2.32}$$

In (2.31), $\overline{Z}_i$ is the mean of $p(\vec{Z})$ in the i-th coordinate. In the Z variables, with (2.31) and (2.32), the Gaussian estimator for the predictability splits into a sum of one dimensional principal predictability factors

$$P(p, \Pi_0) \geq P(p_G^*, \Pi_0) = \sum_{i=1}^M P(p_i^*, \Pi_{0,i}) = \sum_{i=1}^M \frac{1}{2}[\ln \gamma_i^{-1} + \gamma_i - 1] + \sum_{i=1}^M \frac{1}{2}[\overline{Z}_i^2]. \tag{2.33}$$

Clearly, the first term in the summation in (2.33) is the relative entropy contribution to the dispersion from the one dimensional Gaussian distribution p_i^* in (2.31) while the second term arises from the contribution of p_i^* to the signal. Thus, these variables obviously can be organized into groups with higher predictive information content in the dispersion, the signal or the total combination depending on the nature of the application. These variables are called the *Principal Predictability Components*. The change of coordinates in (2.30) will be utilized in Section 4 to get improved estimators beyond Proposition 2.2 for multi-variable distributions in terms of mutual information of joint distributions.

2.4 The Relative Entropy and the Entropy Difference for Quantifying Predictive Information Content

For a normalized Gaussian climate Π_0 and a Gaussian prediction, p_G, Schneider and Griffies (1999) advocate the entropy difference in (2.26)

$$\Delta\mathcal{S}(p_G, \Pi_0) = \mathcal{S}(\Pi_0) - \mathcal{S}(p_G) = \ln(\det(\mathrm{Cov} p_G)^{-1/2}) \tag{2.34}$$

as a measure of predictive information under the tacit assumption that distributions p_G of interest for prediction satisfies $\Delta\mathcal{S}(p_G, \Pi_0) \geq 0$. First, consider the special case of (2.34) for a one-dimensional probability distribution for a quantity with variance for p_G given by σ^2. Clearly, from (2.34) we have

$$\Delta\mathcal{S}(p_G, \Pi_0) < 0 \quad \text{if and only if} \quad \sigma^2 > 1. \tag{2.35}$$

Thus, the entropy difference does not measure information content for distributions with variance $\sigma^2 > 1$. On the other hand, the fact that the variance of a predicted quantity at a short fixed time, such as the temperature at a fixed location, is predicted to be higher than that of the historical climate record has obvious information content ignored in (2.34). With the mathematical property in (1.8), the relative entropy predictive measure always assigns non-trivial information content to these events. Furthermore, as mentioned earlier in Section 2.1, the entropy difference in (2.34) cannot measure the information content expressed in the fact that the mean of the prediction density, p_G, can be non-zero and different from that of Π_0 while the signal contribution to the relative entropy measure of predictive information content developed in Section 2.1 does this precisely. Finally, as developed in 2.1, the relative entropy is invariant under a general change of coordinates while the entropy difference in (2.34) is only unchanged by linear transformation. Quite often quantities such as humidity can be complex nonlinear functions of other variables in the atmosphere/ocean system so a measure of information content which is unchanged by nonlinear transformations is clearly superior to one which is not. For all of these three reasons, we claim that even for Gaussian predictions and climate, the relative entropy measure is superior to the entropy difference in quantifying predictive information content. For general Markov processes such as a finite Markov chain with a non-uniform invariant measure, there are additional compelling dynamic reasons to favor the relative entropy over entropy difference related to the decrease of information as statistical equilibrium is approached (Cover and Thomas (1991)).

3 Explicit Estimators and the Signal-Dispersion Decomposition for Non-Gaussian Climate Variables

Here we generalize the results in Proposition 2.1 and Proposition 2.2 to non-Gaussian climate variables. Concrete applications of the material developed in this section below to non-Gaussian climate variables in illustrative models chaotic dynamical systems is presented elsewhere (Cai, Kleeman, and Majda, 2002 A), B)). For simplicity in exposition, we assume scalar probability distributions in this section and use the notation, $p_c(\lambda)$ for the scalar climate distribution. We do this only for simplicity in writing. All the results presented below extend immediately to multi-variate distributions with more complex notation.

We assume a given climate distribution $p_c(\lambda)$, for a scalar variable which is centered so that it has zero mean,

$$\int \lambda p_c(\lambda) d\lambda = 0. \tag{3.1}$$

We consider the *pdf* from an ensemble prediction, $p_d(\lambda)$. As discussed in the introduction, it is a reasonable strategy in practice to measure some moments of the ensemble prediction distribution, $p_d(\lambda)$,

$$\begin{aligned} &\text{A)} \quad \bar{\lambda}_d = M_1 = \int \lambda p_d(\lambda) d\lambda \\ &\text{B)} \qquad M_j = \int (\lambda - \bar{\lambda}_d)^j \, p_d(\lambda) d\lambda, \qquad 2 \le j \le 2L. \end{aligned} \tag{3.2}$$

For retaining statistical significance in an ensemble prediction, one usually measures only the first two or four moments in practice so that $L = 1$ or $L = 2$ in (3.2) B).

3.1 Theory for Explicit Estimators for Predictability

Here we generalize the discussion in Section 2.2 to non-Gaussian climates. With the moment information from the ensemble prediction in (3.2), it is natural to define p_d^* as the probability measure with the least bias which retains the information in (3.2). To do this, we introduce the set of probability measures which satisfy the $2L$ constraints in (3.2), PM_{2L}. Since $-p \ln\left(\frac{p}{p_c}\right)$ is a concave function of p, we define p_d^* via the maximum entropy principle (Lasota and Mackey, 1994)

$$\mathcal{S}(p_d^*, p_c) = \max_{p_d \in PM_{2L}} \mathcal{S}(p_d, p_c). \tag{3.3}$$

The usual Lagrange multiplier calculation yields the explicit formula for p_d^*,

$$\text{A)} \quad -\ln\left(\frac{p_d^*}{p_c}\right) = \sum_{\substack{j=0\\ j\neq 1}}^{2L} \alpha_j(\lambda - \bar{\lambda}_d)^j + \alpha_1\lambda \tag{3.4}$$

where

B) $\alpha_j, 0 \leq j \leq 2L$ are the Lagrange multipliers for the $2L$ constraints.

With (3.2) and (3.3), p_d^* *provides a rigorous predictability estimator for the ensemble prediction*

$$\mathcal{P}(p_d^*, p_c) \leq \mathcal{P}(p_d, p_c). \tag{3.5}$$

Furthermore, with (3.2) and (3.4), there is an *explicit formula for the Predictability Estimate*,

$$\mathcal{P}(p_d^*, p_c) = -\sum_{j=2}^{2L} \int p_d^* \alpha_j(\lambda - \bar{\lambda}_d)^j - (\alpha_0 + \bar{\lambda}_d\alpha_1). \tag{3.6}$$

Next, assume that we have ensemble predictability estimators, p_{d,L_1}^*, p_{d,L_2}^* where p_{d,L_1}^* involves $2L_1$ constraints and p_{d,L_2}^* involves $2L_2$ constraints with $L_1 < L_2$. By the definition in (3.3), we have

$$\mathcal{S}(p_{d,L_1}^*, p_c) \geq \mathcal{S}(p_{d,L_2}^*, p_c) \tag{3.7}$$

which yields the *Predictability Estimator Principle*:

$$\mathcal{P}(p_{d,L_1}^*, p_c) \leq \mathcal{P}(p_{d,L_2}^*, p_c) \leq \mathcal{P}(p_d, p_c) \tag{3.8}$$

Clearly this theory generalizes the one in Section 2.2 to non-Gaussian climate variables

Remark. We can anticipate that the above predictability estimators will be quite useful in problems where the ensemble prediction distribution is **not** highly intermittent; if the tail of the ensemble prediction distribution matters a lot, then the above predictors will probably perform quite poorly. See Cai, Kleeman, Majda (2002 A), B)) for discussion of this issue.

3.2 The Generalized Signal and Dispersion Decomposition

Here we define useful signal and dispersion contributions for a given ensemble prediction generalizing the case of a Gaussian climate in Section 2.1. To begin our discussion, we examine the decomposition for the predictability estimate in (3.6) which we recall here,

$$\mathcal{P}(p_d^*, p_c) = -\sum_{j=2}^{2L} \int p_d^* \alpha_j(\lambda - \bar{\lambda}_d)^j - (\alpha_0 + \bar{\lambda}_d\alpha_1). \tag{3.9}$$

The *first term* in (3.9) *contributes,* $\mathcal{D}_1$ *directly to* the *dispersion* by construction since

$$\mathcal{D}_1 = \int p_d^* \left(-\sum_{j=1}^{2L} \alpha_j (\lambda - \bar{\lambda}_d)^j \right) \tag{3.10}$$

so that $\mathcal{D}_1$ reflects the contributions from the variance, skewness, kurtosis, etc. of the prediction ensemble centered about the ensemble mean. Similarly, the last term in (3.9) makes a direct contribution, $\mathcal{S}_1$, to the signal,

$$\mathcal{S}_1 = -\bar{\lambda}_d \alpha_1 \tag{3.11}$$

since the mean enters explicitly in $\mathcal{S}_1$. The subtle terms in the signal-dispersion contribution involve a splitting of α_0 into signal and dispersion pieces.

To discuss the decomposition of α_0, we make a natural simplifying assumption on the climate distribution, p_c. We assume that the climate distribution, p_c, also arises from a maximum entropy principle expressing the measure with least bias given information on moments, i.e. p_c satisfies

$$\mathcal{S}(p_c) = \max_{p \in PM_L^c} \mathcal{S}(p) \tag{3.12}$$

where

$$\mathcal{S}(p) = -\int p \ln p. \tag{3.13}$$

Here PM_{2L}^c is the set of probability measures satisfying the imposed constraints on climate moments so that

$$\begin{aligned} &p \in PM_{2L}^c \quad \text{means} \\ &\int \lambda p = 0 \\ &\int \lambda^j p = M_j^c, \qquad 2 \leq j \leq 2L \end{aligned} \tag{3.14}$$

with M_j^c, the prescribed measured values of the j-th moment in the climate. The climate measure, p_c, determined from (3.12) has the least bias given the statistical measurements in (3.14). Thus, (3.12) is the most natural way to do climate measurements given the behavior of climate moments in (3.14). Also, clearly for $L = 1$ this construction reduces to the case for Gaussian climate variables discussed in Section 2.

The standard Lagrange multiplier calculation for (3.12), (3.14) yields the formula

$$p_c = e^{-\sum_{j=1}^{2L} \alpha_j^c \lambda^j} e^{-\alpha_0^c} \tag{3.15}$$

where α_j^c are the Lagrange multipliers for the climate moments guaranteeing the constraints in (3.14) and

$$e^{\alpha_0^c} = \int e^{-\sum_{j=1}^{2L} \alpha_j^c \lambda^j} d\lambda. \tag{3.16}$$

For the special climate distributions defined in (3.15), it follows from (3.4) A) and (3.15) that the constant $-\alpha_0$ in (3.9) is given by

$$-\alpha_0 = -\ln\left(\int e^{-\left(\sum_{j=2}^{2L}\alpha_j(\lambda-\bar{\lambda}_d)^j+\alpha_1\lambda+\sum_{j=1}^{2L}\alpha_j^c\lambda^j+\alpha_0^c\right)}\right). \tag{3.17}$$

Take the polynomial of degree $2L$ in the exponent of the integrand and trivially explicitly recenter this polynomial about $\bar{\lambda}_d$ as follows,

$$\begin{aligned}\sum_{j=2}^{2L}\alpha_j(\lambda-\bar{\lambda}_d)^j+\alpha_1\lambda+\sum_{j=1}^{2L}\alpha_j^c\lambda^j &= \sum_{j=1}^{2L}\alpha_j^{d,c}(\lambda-\bar{\lambda}_d)^j-\bar{\alpha}_0^{d,c} \\ &\underset{\text{Def}}{\equiv} \phi^{d,c}(\lambda-\bar{\lambda}_d)-\bar{\alpha}_0^{d,c}.\end{aligned} \tag{3.18}$$

With (3.18) and (3.17),

$$-\alpha_0 = \alpha_0^c + \bar{\alpha}^{d,c} - \ln\int e^{-\phi^{d,c}(\lambda-\bar{\lambda}_d)}d\lambda. \tag{3.19}$$

Which parts of (3.19) contribute to the signal and the dispersion? Clearly, the value of

$$\ln\int e^{-\phi^{d,c}(\lambda-\bar{\lambda}_d)}d\lambda = \ln\int e^{-\phi^{d,c}(\lambda-\tau)}d\lambda \tag{3.20}$$

for any shift τ so that this term contributes to the dispersion although $\phi^{d,c}$ has Lagrange multiplier coefficients itself that are functions of $\bar{\lambda}$; also, α_0^c defined in (3.16) involves purely the climate alone and does not depend on $\bar{\lambda}_d$ so this term cannot contribute to the signal and instead normalizes the dispersion. On the other hand, from (3.18), $\bar{\alpha}_0^{d,c}$ clearly is a function of $\bar{\lambda}_d$ and should be grouped with the signal. Thus, respective contributions to the signal and dispersion from $-\alpha_0$ in (3.17) are defined by

$$-\alpha_0 = \mathcal{S}_2 + \mathcal{D}_2 \tag{3.21A}$$

where

$$\mathcal{S}_2 = \bar{\alpha}_0^{d,c} \tag{3.21B}$$

and

$$\mathcal{D}_2 = \alpha_0^c - \ln\int e^{-\phi^{d,c}(\lambda-\bar{\lambda}_d)}d\lambda \text{ with } \phi^{d,c} \text{ defined in (3.18).} \tag{3.21C}$$

By combining (3.9) to (3.21) we obtain the following

Proposition 3.1 (Generalized Signal-Dispersion Decomposition). *Assume that the climate distribution, p has the special form in* (3.15). *Then the predictability estimate $\mathcal{P}(p_d^*, p_c)$ admits the explicit decomposition into signal and dispersion,*

$$\mathcal{P}(p_d^*, p_c) = \mathcal{S} + \mathcal{D} \tag{3.22}$$

where the signal $\mathcal{S}$ is given by

$$\mathcal{S} = \bar{\alpha}_0^{d,c} - \bar{\lambda}_d \alpha_1 \tag{3.23}$$

with $\bar{\alpha}_0^{d,c}$ defined in (3.18) *and the dispersion is given by*

$$\mathcal{D} = \mathcal{D}_1 + \mathcal{D}_2 \tag{3.24}$$

where $\mathcal{D}_1$ is given in (3.10) *and $\mathcal{D}_2$ is given in* (3.21C).

It is a simple exercise for the reader to show that the decomposition in Proposition 2.1 is recovered for a Gaussian climate.

4 Practical Predictability Bounds for Many Degrees of Freedom

Here we address the practical issue in (1.12) C). How can one practically compute or approximate the relative entropy, $\mathcal{P}(p, \Pi)$, for a complex dynamical system with many degrees of freedom and a subset of variables of dimension N with $N \gg 1$? Below, we show how to obtain a rigorous lower bound on the information content of an ensemble prediction through a sum of

A) N one-dimensional relative entropies

B) N two-dimensional mutual informations of suitable marginal distributions.

The practical significance is that the sum of N one and two dimensional relative and mutual entropies can be evaluated rapidly and efficiently for $N \gg 1$ while the direct evaluation of the relative entropy in (1.7) as an N-dimensional integral is prohibitively expensive or practically impossible. Yet, we establish below with full mathematical rigor that a sum of such quantities provides a lower bound on the predictive information content.

As discussed in the introduction, the natural predictability measure is the relative entropy,

$$\mathcal{P}(p, \Pi) = \int_{R^N} p \ln\left(\frac{p}{\Pi}\right) d\vec{x}. \tag{4.1}$$

Here we make the additional assumption that the variables $\vec{x}$ are chosen conveniently so that the probability distribution, $\Pi(\vec{x})$, with respect to these variables has the *Factored Form*:

$$\Pi(\vec{x}) = \prod_{j=1}^{N} \Pi_j(x_j). \tag{4.2}$$

This occurs in many applications and in particular for the Gaussian case.

Next for a factored climate measure, $\Pi(\vec{x})$, we show that $\mathcal{P}(p, \Pi)$ splits into a sum of two interesting relative entropy measures. First, we introduce the *marginal probability* distributions, $p_{\vec{j}}(\vec{x}_{\vec{j}})$ for the $|\vec{j}|$ variables $\vec{x}_j$ where $\vec{j} = (j_1, \ldots, j_{|\vec{j}|})$ and without loss of generality, $j_1 < j_2 < \cdots < j_{|\vec{j}|}$. The $p_{\vec{j}}$ are defined by

$$p_{\vec{j}}(\vec{x}_{\vec{j}}) = \int_{R^{N-|\vec{j}|}} p(x_1, \ldots, x_N) \prod_{\substack{L \neq j_i \\ 1 \leq i \leq |\vec{j}|}} dx_L. \tag{4.3}$$

Below we omit the parentheses in denoting $p_{\vec{j}}(\vec{x}_{\vec{j}})$. Next, we expand the relative entropy measure via

$$\begin{aligned} \mathcal{P}(p, \Pi) &= \int_{R^N} p \ln p - \int_{R^N} p \ln \Pi \\ &= \int_{R^N} p \ln p - \sum_{j=1}^{N} \int_{R^1} p_j(x_j) \ln \Pi_j(x_j)\, dx_j \\ &= \int_{R^N} p \left(\ln p - \sum_{j=1}^{N} \ln p_j \right) d\vec{x} + \sum_{j=1}^{N} \int_{R^1} p_j (\ln p_j - \ln \Pi_j(x_j)) dx_j \\ &= \mathcal{P}\left(p, \prod_{j=1}^{N} \otimes p_j \right) + \sum_{j=1}^{N} \mathcal{P}(p_j, \Pi_j). \end{aligned} \tag{4.4}$$

We record the identity in (4.4) as the following *Predictability Measure Decomposition*:

Proposition 4.1. *For a factored climate satisfying* (4.2)*, the relative entropy predictability measure exactly splits into the decomposition,*

$$\mathcal{P}(p, \Pi) = \mathcal{P}\left(p, \prod_{j=1}^{N} \otimes p_j \right) + \sum_{j=1}^{N} \mathcal{P}(p_j, \Pi_j). \tag{4.5}$$

The importance of (4.5) is that the term

$$\sum_{j=1}^{N} \mathcal{P}(p_j, \Pi_j) \text{ is a sum of one-dimensional predictability contributions} \tag{4.6B}$$

relative to the climate while

$$\mathcal{P}\left(p, \prod_{j=1}^{N} \otimes p_j\right) \text{ involves only the prediction probability density alone and} \tag{4.6B}$$

the extent to which it factors into a product.

Since $\mathcal{P}\left(p, \prod_{j=1}^{N} \otimes p_j\right) \geq 0$ always we have the *Predictability Bound for Factored Climate*

$$\mathcal{P}(p, \Pi) \geq \sum_{j=1}^{N} \mathcal{P}(p_j, \Pi_j) \tag{4.7}$$

i.e. the sum of the $1-D$ marginals always bound the predictive information content. While practically this is less useful, we also note that by the same reasoning

$$\mathcal{P}(p, \Pi) \geq \mathcal{P}\left(p, \prod_{j=1}^{N} \otimes p_j\right). \tag{4.8}$$

Of course, the bound in (4.8) completely ignores the information in the climate record. However, at very short times in a prediction, this other term could be very important.

For the extreme special case with $N = 2$,

$$\mathcal{P}(p, p_1 \otimes p_2) \text{ is the } \textit{mutual information} \tag{4.9}$$

i.e. a measure of the information content in a probability distribution beyond its factors with $\mathcal{P}(p, p_1 \otimes p_2) = 0$ only for x_1, x_2 independent random variables (Cover and Thomas, 1991). Clearly we have the expansion in terms of the entropy in (4.4),

$$\mathcal{P}(p, p_1 \otimes p_2) = -[\mathcal{S}(p) - (\mathcal{S}(p_1) + \mathcal{S}(p_2))]. \tag{4.10}$$

Proposition 4.1 shows that for factored climates, one needs to obtain useful lower bounds on the relative entropy distribution

$$\mathcal{P}\left(p, \prod_{j=1}^{N} \otimes p_j\right)$$

in addition to estimating the one-dimensional entropies in order to obtain practical estimates for the information content. Next, we show how to obtain lower bounds on this quantity in general in terms of N-two dimensional mutual informations. Our discussion relies on two well-known mathematical facts (see Cover and Thomas, 1991):

Taking Marginals Always Reduces Relative Information

$$\mathcal{P}(p,\Pi) \geq \mathcal{P}(p_{\vec{j}},\Pi_{\vec{j}})$$

for any marginal distributions defined by $\vec{j}$.

SubAdditive Information Bound

Consider an arbitrary marginal, $p_{\vec{j}}(\vec{x}_j)$ and split $\vec{j} = \vec{j}_1 \cup \vec{j}_2 \cup \vec{j}_3$ as a disjoint union of three sets. Define

p_{12} is marginal defined w.r.t. vector $\vec{j}_1 \cup \vec{j}_2$

p_{23} is marginal defined w.r.t. vector $\vec{j}_2 \cup \vec{j}_3$

p_2 is marginal defined by $\vec{j}_2$

p_{123} is marginal, $p_{\vec{j}}(\vec{x}_j)$, defined by $\vec{j} = \vec{j}_1 \cup \vec{j}_2 \cup \vec{j}_3$.

We have the *SubAdditive Information Bound*

$$\mathcal{S}(p_{123}) + \mathcal{S}(p_2) \leq \mathcal{S}(p_{12}) + \mathcal{S}(p_{23})$$

These two principles are utilized in obtaining lower bounds as follows; Subadditivity yields

$$\begin{aligned}
\mathcal{P}\left(p, \bigotimes_{i=1}^{N} p_i\right) &= -\mathcal{S}(p) + \sum_{i=1}^{N} \mathcal{S}(p_i) \\
&\geq \mathcal{S}(p_{3,\ldots,N}) - \mathcal{S}(p_{1,3,\ldots,N}) - \mathcal{S}(p_{2,3,\ldots,N}) + \sum_{i=1}^{N} \mathcal{S}(p_i) \\
&= [-\mathcal{S}(p_{1,3,\ldots,N}) + \mathcal{S}(p_{3,\ldots,N}) + \mathcal{S}(p_1)] + [-\mathcal{S}(p_{2,3,\ldots,N}) + \mathcal{S}(p_{3,\ldots,N}) + \mathcal{S}(p_2)] \\
&\quad + \left[-\mathcal{S}(p_{3,\ldots,N}) + \sum_{i=3}^{N} \mathcal{S}(p_i)\right] \equiv \mathcal{P}(p_{1,3,\ldots,N}, p_1 \otimes p_{3,\ldots,N}) \qquad (4.11) \\
&\quad + \mathcal{P}(p_{2,3,\ldots,N}, p_2 \otimes p_{3,\ldots,N}) + \mathcal{P}\left(p_{3,\ldots,N}, \bigotimes_{i=3}^{N} p_i\right).
\end{aligned}$$

Next the lower bound for marginals yields

$$\begin{aligned}
&\text{A)}\quad \mathcal{P}(p_{1,3,\ldots,N}, p_1 \otimes p_{3,\ldots,N}) \geq \mathcal{P}(p_{1,3}, p_1 \otimes p_3) \\
&\text{B)}\quad \mathcal{P}(p_{2,3,\ldots,N}, p_2 \otimes p_{3,\ldots,N}) \geq \mathcal{P}(p_{2,3}, p_2 \otimes p_3).
\end{aligned} \qquad (4.12)$$

Combining (4.11) and (4.12) yields

Lemma 4.1. *For general N*

$$\mathcal{P}\left(p, \bigotimes_{i=1}^{N} p_i\right) \geq \mathcal{P}(p_{1,3}, p_1 \otimes p_3) + \mathcal{P}(p_{2,3}, p_2 \otimes p_3) + \mathcal{P}\left(p_{3,\ldots,N}, \bigotimes_{i=3}^{N} p_i\right). \qquad (4.13)$$

Remark. The advantage of (4.13) is that it can be iterated and applied successively to $\mathcal{P}\left(p_{3,\ldots,N}, \bigotimes_{i=3}^{N} p_i\right)$ to yield bounds in terms of the mutual information i.e. applying the lemma on $p_{3,\ldots,N}$ yields

$$\mathcal{P}\left(p_{3,\ldots,N}, \bigotimes_{i=3}^{N} p_i\right) \geq \mathcal{P}(p_{3,5}, p_3 \otimes p_5) + \mathcal{P}(p_{4,5}, p_4 \otimes p_5) + \mathcal{P}\left(p_{5,\ldots,N}, \bigotimes_{i=5}^{N} p_i\right). \tag{4.14}$$

Combining Proposition 4.1 and continuing these calculations we obtain a *Lower Bound for Predictability via 1-Variable Relative Entropy and 2-Variable Mutual Information for a Factored Climate*.

Proposition 4.2. *Consider a general predictability distribution for general $p(\vec{x}_N)$ (for any permutation of $x_1 \ldots x_N$)*
1) If N is odd

$$\begin{aligned}\mathcal{P}(p, \Pi) \geq & \sum_{j=0}^{\frac{N-3}{2}} [\mathcal{P}(p_{2j+1,2j+3}, p_{2j+1} \otimes p_{2j+3}) + \mathcal{P}(p_{2j+2,2j+3}, p_{2j+2} \otimes p_{2j+3})] \\ & + \sum_{j=1}^{N} \mathcal{P}(p_j, \Pi_j).\end{aligned} \tag{4.15}$$

2) If N is even

$$\begin{aligned}\mathcal{P}(p, \Pi) \geq & \sum_{j=0}^{\frac{N-4}{2}} [\mathcal{P}(p_{2j+1,2j+3}, p_{2j+1} \otimes p_{2j+3}) + \mathcal{P}(p_{2j+2,2j+3}, p_{2j+2} \otimes p_{2j+3})] \\ & + \mathcal{P}\left(p_{N-1,N}, p_{N-1} \otimes p_N\right) + \sum_{j=1}^{N} \mathcal{P}(p_j, \Pi_j).\end{aligned} \tag{4.16}$$

Proposition 4.2 achieves the practical goal of estimating the information content rigorously through N one-dimensional relative entropies and N two-dimensional mutual informations. Practical applications of these ideas will be presented elsewhere to models of interest in the geosciences with many degrees of freedom.

We end this section with a simple remark for a Gaussian climate, Π_0. If we utilize the special variables ordering predictive information content from Section 2.3 above, then for the Gaussian predictability estimator, p_G^*, it follows easily from (2.33) and Proposition 2.1 that

$$\sum_{i=1}^{N} \mathcal{P}(p_i, \Pi_i) \geq \sum_{i=1}^{N} \mathcal{P}(p_{G,i}^*, \Pi_i) \equiv \mathcal{P}(p_G^*, \Pi_0).$$

Thus, when the special basis ordering information content and depending on p is utilized for a Gaussian climate, the one-dimensional relative entropies already have higher information content than the simplest Gaussian estimator and any improved estimates of these as well as any use of the mutual information automatically increases our knowledge of the information content in the prediction.

4.1 The Signal-Dispersion Decomposition for Many Variables

Finally we make a remark which allows us to generalize the Signal-Dispersion decomposition developed in Section 2.1 and Section 3 for a multi-variable probability distribution with a factored climate. First, observe from (4.4) that

$$\mathcal{P}\left(p, \prod_{j=1}^{N} \otimes p_j\right) = -\mathcal{S}(p) + \sum_{j=1}^{N} \mathcal{S}(p_j)$$

and also that the entropy is translation invariant, $\mathcal{S}(p) = \mathcal{S}(p_\tau)$ for any shift τ. Thus, necessarily $\mathcal{P}\left(p, \prod_{j=1}^{N} \otimes p_j\right)$ contributes only to the dispersion of the signal-dispersion decomposition. Under the assumptions of Section 3 on Π_j Proposition 3.1 guarantees that the one-dimensional distributions $\mathcal{P}\left(p_j, \Pi_j\right)$ have the signal-dispersion decomposition

$$\mathcal{P}\left(p_j, \Pi_j\right) = \mathcal{D}_j + \mathcal{S}_j \ .$$

Combining the above two remarks, we obtain the generalized *Signal-Disperson Decomposition*

Proposition 4.3. *For a factored climate with Π_j determined as in Section 3,*

$$\mathcal{P}\left(p, \Pi\right) = \mathcal{S} + \mathcal{D}$$

where the signal is the sum of the univariate signals

$$\mathcal{S} = \sum_{j=1}^{N} \mathcal{S}_j$$

while

$$\mathcal{D} = \sum_{j=1}^{N} \mathcal{D}_j + \mathcal{P}\left(p, \prod_{j=1}^{N} \otimes p_j\right)$$

with $\mathcal{D}_j$ the univariate dispersions where $\mathcal{S}_j, \mathcal{D}_j$ are determined in Proposition 3.1.

Acknowledgment: Andrew Majda acknowledges generous support from the Office of Naval Research and the National Science Foundation for this research.

References

Cai, D., Kleeman, R. and Majda, A.J., 2002a, "Test models for quantifying predictability, I: Simple chaotic maps", *in preparation.*

Cai, D., Kleeman, R. and Majda, A.J., 2002b, "Test models for quantifying predictability, II: The Kuramoto-Sivashinsky equation", *in preparation.*

Cover, T.M. and Thomas, J.A., 1991, *Elements of Information Theory*, New York: Wiley.

Kleeman, R., 2002, "Measuring dynamical prediction utility using relative entropy", *J. Atmos. Sci.* **59**, 2057–2072.

Kleeman, R., Majda, A.J. and Timofeyev, I., 2002, "Quantifying predictability in a model with statistical features of the atmosphere", submitted to *Physica D* in November 2001.

Lasota, A. and Mackey, M., 1994, *Chaos, Fractals, and Noise*, Applied Math. Sciences 97, Springer-Verlag, New York.

Majda, A.J. and Timofeyev, I., 2000, "Remarkable statistical behavior for truncated Burgers–Hopf dynamics", *Proc. Nat. Acad. Sci.* **97**(23), 12413–12417.

Majda, A.J., Timofeyev, I. and Vanden Eijinden, E., 2001, "A mathematical framework for stochastic climate model", *Comm. Pure Appl. Math.,* Vol. **54**, Issue 8, 891–974.

Majda, A.J. and Wang, X., 2002, *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*, Based on Lectures by A.J. Majda at the Courant Institute (in final preparation for Cambridge Univ. Press).

Olla, S., Varadhan, S.R.S., and Yau, H.T., 1993, "Hydrodynamic limit for a Hamiltonian system with weak noise", *Comm. Math. Phys.* **155**, 523–560.

Schneider, T. and Griffies , 1991, "A conceptual framework for predictability studies", *J. Clim.* **12**, 3133–3155.

Schutte, Ch., Fisher, A., Huisinga, W., and Deuflhard, P., 1999, "A direct approach to conformation dynamics based on hybrid Monte Carlo", *J. Comp. Phys.* **151**, 146–168.

Toth, Z., 1991, "Circulation patterns in phase space: A multinormal distribution? *Mon. Wea. Rev.* **119**, 1501–1511.

Varadhan, S.R.S., 1984, "Large deviations and applications", CBMS-NSF Regional Conference Series, Number 46, *SIAM.*

Varadhan, S.R.S., 1985, "Large deviations and applications", *Expo. Math.* **3**, 251–272.

Yau, H.T., 1991, "Relative entropy and hydrodynamics of Ginzburg–Landau models", *Lett. Math. Phys.* **22**, 63–80.