Predictability and Information Theory

Part I: Measures of Predictability

Timothy DelSole

Center for Ocean-Land-Atmosphere Studies

George Mason University

Draft Version: May 2, 2003

Abstract

The connection between predictability and information theory is discussed with the aim of clarifying the various metrics which have been proposed in the literature. This discussion begins with the following definition of predictability: *an event is predictable if the posterior distribution, after a set of observations becomes available, differs in any way from the prior distribution which is known before observations become available.* In accordance with this definition, predictability is measured by the difference between two distributions. Three such measures have been proposed in the literature: predictive information, relative entropy, and mutual information. All three metrics have attractive properties for defining predictability, including the fact that they are invariant with respect to nonsingular linear transformations, decrease monotonically in stationary Markov systems in some sense, and are easily decomposed into components which optimize them (in certain cases). Relative entropy and predictive information have the same average value, which in turn is precisely equal to the mutual information. Thus, mutual information is distinguished from the other two quantities in that it represents the average predictability over all initial conditions. Relative entropy is distinguished from predictive information in that it vanishes if and only if the two distributions are identical. This difference can be of fundamental importance in some cases. Predictable component analysis, a technique for decomposing predictive information into components that optimize this quantity, also decomposes relative entropy when the prior and posterior distributions have the same mean. Optimization of mutual information is shown to be equivalent to canonical correlation analysis when the variables are Gaussian; the canonical patterns can be interpreted as components with optimum average predictability.

1 Introduction

Recently, several papers have discussed the connection between information theory and the predictability of weather and climate (Leung and North 1990; Schneider and Griffies 1999; Kleeman 2002). These papers present different definitions of predictability, and present different metrics for measuring predictability. The reasons for choosing different metrics, and the connection between the metrics, are not always explained in these papers. In addition, the literature on this subject has been confined almost exclusively to the case of perfect model scenarios. The goal of this paper is to clarify the relation between predictability and the various concepts in information theory that have been proposed. This paper will concentrate on theoretical issues; applications can be found in the above cited papers. In part II of this paper, we discuss methods for dealing with imperfect models which were suggested by Schneider and Griffies (1999).

2 Basic Concepts in Predictability Theory

The need for a new framework for predictability studies can be appreciated from the fact that there exist problems with defining predictability. For example, all weather predictability is said to be lost when the error in predicting the weather is comparable to the difference between two randomly chosen states of the system. A problem with this definition, as a universal one, emerges as soon as one attempts to define the predictability of climate change. In climate change problems, we are not interested in how close the true state is predicted, but how the statistics change due to some perturbation. Forecast error is a poor measure of such change. This definition also is problematic when the

forecast is probabilistic. In such cases, “error” is not defined naturally. If error is defined as the mean square difference between the true state and ensemble mean, then this error oscillates irregularly with lead time. Such fluctuations would imply that the “limit of predictability” also fluctuates in time, perhaps becoming small at irregular intervals, giving the appearance of predictability (by this definition) when in fact the occurrence of small error is itself a random process which cannot be predicted. A second definition could be based on the forecast error averaged over all initial conditions. This definition would eliminate the above fluctuations, but would leave the predictability of a single event undefined. In addition to these problems, no agreement exists on how to combine different variables to obtain a single measure of error or predictability– e.g., how many degrees Celsius of temperature is equivalent to one meter per second of wind velocity. Also, mean square error depends on the basis set in which the data is represented, which is unsatisfactory since the basis set is arbitrary. For these reasons, a universal definition of predictability is difficult to formulate in terms of forecast error.

To formulate a single definition of predictability that encompasses both weather and climate, it is necessary to review some basic concepts of predictability theory. The following represents a theoretical framework which seems consistent with the literature. This discussion assumes a perfect model scenario, by which we mean that the governing equations and all required probability distributions of the system are known.

Predictability is the study of the extent to which events can be predicted. The foundations of predictability theory were laid by Lorenz (1963, 1965, 1969) and Epstein

(1969). The basic idea is that the initial state of a system is not known exactly and therefore is represented most properly by a probability distribution function (pdf), which can be interpreted as a density of possible states in phase space. The distribution evolves according to Liouville's equation, which is derived the same way as the mass continuity equation, under the principle that no ensemble members are created or destroyed (i.e., the total amount of probability always equals unity). A generalized version of Liouville's equation, called the Fokker-Planck equation, applies if the dynamical model is stochastic– that is, if the future state is not uniquely determined by the initial state, as would be the case if certain physical processes were represented by random processes.

The probability distribution of a system changes discontinuously after the system is observed. This event is called "collapse of the wavefunction" in quantum mechanics. The resulting distribution, called the *conditional* distribution, gives the distribution of states given knowledge about the system gained through observation. Typical examples of observations include the state itself, boundary condition, and/or external forcing. The distribution before observations become available is called the *prior* distribution, while the distribution after observations become available is called the *posterior* distribution. (Note that this terminology differs from that of Schneider and Griffies 1999). Let the variable at time t be x_t, and let the set of realizations of the observations up to time t be o_t. Then we denote the distribution before observations become available by $p(x_t)$, and the distribution conditioned on all observations up to time t by $p(x_t|o_t)$. Both distributions are related in classical probability theory through Bayes' theorem. Statistical procedures for constructing $p(x_t|o_t)$ were developed by Wiener and Kalman and now have been

generalized quite extensively (Jazwinski 1970). We assume in this paper that this procedure has been performed and $p(x_t|o_t)$ is known.

In predictability theory, the variable X is predicted at a future time t + τ, where τ is a positive lead time. The most complete description of a prediction is the distribution of the variable conditioned on the observation, $p(x_{t+\tau}|o_t)$. This distribution, called the posterior distribution, can be computed from the classical formula

$$p(x_{t+\tau}\,|\,o_t) = \sum_{x_t} p(x_{t+\tau}\,|\,x_t)\; p(x_t\,|\,o_t)\,, \qquad \textbf{(1)}$$

where $p(x_{t+\tau}\,|\,x_t\,,\,o_t) = p(x_{t+\tau}\,|\,x_t)$ has been invoked, owing to the fact that $x_{t+\tau}$ depends only on x_t when x_t is known (this assumes that the observation does not disturb the system). The sum is replaced by an appropriate integral if x_t is continuous. Here, discrete forms will be used for convenience of notation. The distribution $p(x_{t+\tau}\,|\,x_t)$ is called a transition probability and is computed from a dynamical or stochastic model. In a perfect model scenario, the posterior distribution $p(x_{t+\tau}|o_t)$ is simply the distribution of the forecast ensemble generated from a set of initial states randomly selected from $p(x_t|o_t)$. For Markov systems, $p(x_{t+\tau}\,|\,o_t)$ satisfies a Fokker-Planck equation with initial condition $p(x_t\,|\,o_t)$. If the system is stationary, then the prior distribution $p(x_{t+\tau})$ is most appropriately identified with the asymptotic forecast; i.e., $p(x_t) = p(x_{t+\tau}) = \lim p(x_{t+\tau}|o_t)$ as $\tau \to \infty$.

If observations do not alter a distribution, then the associated event is said to be unpredictable with respect to those observations. Equivalently, *an event is predictable if*

the posterior distribution, after a set of observations becomes available, differs in any way from the prior distribution which is known before observations become available.

It is crucial to recognize that the above definition defines predictability by the *difference* in probability distributions. Thus, the statement "event *X* is not predictable" is not an absolute statement, rather, it is a relative statement meaning that the observations in question did not modify what was already known about the event relative to the prior distribution. No event is truly unpredictable, since the prior distribution always constitutes a prediction of an "unpredictable" event. Also, the above definition characterizes the universe of all predictable events, without regard to their "usefulness." Clearly, no universal definition of "usefulness" exists, since it depends on geographical, economic, or societal issues that differ from user to user. However, the set of useful predictions must be a subset of all predictable events. Thus, in practice, the above definition of predictability will be supplemented by a more specialized definition to obtain a subset of events that are both predictable and useful.

All loss of predictability occurs when the prior distribution $p(x_{t+\tau})$ equals the posterior distribution $p(x_{t+\tau}|o_t)$. This by definition implies that the variable $x_{t+\tau}$ is statistically independent of the observations o_t. From (1) it follows that necessary conditions for an event to be predictable are that observations should give information about the initial state, and that that initial state should give information about the future.

Another implication of the above definition is that forecasts based on the posterior distribution may be associated with more uncertainty than those based on the prior. That

is, a “predictable event” is not necessarily “more predictable,” in these sense that it can be predicted with less uncertainty relative to a prediction based on the prior. This flexibility allows us to speak about increases in uncertainty as predictable events.

The question arises as to what distribution should define the prior distribution. No unique prior appears to exists– distributions under different conditioning events all constitute different aspects of the total problem. In stationary Markov systems, the most appropriate prior is the unconditional distribution– the distribution of states when no (recent) observation regarding the system is available. For other systems, however, the unconditional distribution is not always an appropriate prior. For instance, in the case of the earth’s climate system, the unconditional distribution of states would include ice ages. But a distribution that includes ice ages would be an inappropriate prior for assessing the predictability of tomorrow’s weather. This suggests that the prior depends on the time scale of interest. In point of fact, the prior need not be associated with real observations at all. For instance, consider the question of whether precipitation at some site is more likely to occur during a certain phase of ENSO. In this case, the phase of ENSO is not a real event, but rather a hypothetical event invented for the purpose of understanding the effect of ENSO on precipitation. Whatever the prior, however, the posterior and prior should differ only by the set of observations used for conditioning.

Lorenz (1975) distinguished two kinds of climate predictability: the first kind refers to the extent to which statistics over a fixed time span change as the beginning and end of the time space advance, while the second kind refers to the extent to which

statistics change in response to changes in climate forcing (the reader should be warned that Lorenz's definitions often are misquoted in the literature). The discussion so far has been restricted to predictability of the first kind. Predictability of the second kind is distinguished by the fact that both the prior and posterior are conditional distributions *of a larger climate distribution*. For example, in the case of global warming, the prior distribution is identified with the climate distribution conditioned on the current greenhouse gas composition, while the posterior distribution is conditioned on the higher greenhouse gas level. If the original concentration is θ_0 and the increased level is θ_1, then the prior would be $p(x_{t+\tau}|\theta_0)$ while the posterior would be $p(x_{t+\tau}|\theta_1)$.

Despite the fundamental difference between predictability of the first and second kinds, both kinds can be sensibly quantified by the difference between the prior and posterior distributions. This definition is natural for predictability of the second kind, since climate change is practically defined by the difference between two distributions. In predictability of the first kind, our definition appears to be unconventional, but still gives sensible results nonetheless. By our definition, weather predictability is measured not by the dispersion of the posterior distribution, as in the paradigm based on forecast error, but by the "distance" between the posterior and stationary distribution to which all forecasts approach. We show in sec. 4 that, for normally distributed, stationary Markov systems, the above definition gives the same limit of predictability as that based on forecast error. Therefore, the above definition appears to provide a single, consistent framework for quantifying different kinds of predictability.

3 Entropy, Relative Entropy, and Mutual Information

While we have defined predictability by the difference between the prior and posterior distributions, how are we to quantify the difference between two distributions? The goal of this section is to show that defining the difference based on information theoretic principles leads to an intuitive methodology.

Consider the prediction of an event. Before the event occurs, we are uncertain as to the outcome of the event. The amount of uncertainty is related in some way to the probability distribution of the event. After the event has been observed, the uncertainty has been removed, and we may say that we have received some *information* through observation. Thus, a decrease in uncertainty corresponds to an increase in information. If two events X and Y are independent, the information h gained when we learn both X and Y should equal the sum of the information gained if X alone were learned, and the information gained if Y alone were learned; i.e., $h(X \text{ and } Y) = h(X) + h(Y)$. It can be shown that the logarithm is the only differentiable, composite function of probability (to within a multiplicative factor) that has this additivity property for independent events. Thus, a measure of the information content of event X with probability $p(x)$ is

$$h(X = x) = \log \frac{1}{p(x)} . \qquad \textbf{(2)}$$

Note how this measure conforms with our intuition about information: if $p(x) = 99.99\%$, then knowledge that event x occurred gives very little information, since it had a high

probability of occurrence to begin with; if $p(x)$ is very small, then knowledge that event x occurred conveys a great deal of information, since x is rare. The average information, weighted by the probability of each event, is called the entropy and is given by

$$H(X) = \sum_x p(x)\, h(p(x)) = -\sum_x p(x) \log p(x) . \qquad \textbf{(3)}$$

There exist many excellent reviews of entropy in the literature (Shannon 1948; Goldman 1953; Reza 1961; Cover and Thomas 1991). These reviews give numerous compelling arguments demonstrating that entropy arises as a natural and fundamental measure of uncertainty in communication theory, data compression, gambling, computational complexity, statistics, and statistical mechanics. We will not reproduce those arguments here, as we could hardly do them justice. Nevertheless, if we want a measure of uncertainty to have properties consistent with the meaning of the concept, then it is hard to escape the fact that we are *compelled* to use entropy as a measure of uncertainty.

There are several ways in which to measure the difference between the posterior and prior distributions in the context of information theory. One approach is to measure predictability by the entropy difference between the prior and posterior distributions:

$$\begin{aligned} P_O &= H(X_{t+\tau}) - H(X_{t+\tau} \mid O = o_t) \\ &= -\sum_{x_{t+\tau}} p(x_{t+\tau}) \log p(x_{t+\tau}) + \sum_{x_{t+\tau}} p(x_{t+\tau} \mid o_t) \log p(x_{t+\tau} \mid o_t) . \end{aligned} \qquad \textbf{(4)}$$

Schneider and Griffies (1999) introduce a related measure called predictive information, defined as the entropy of the prior distribution minus that of the *forecast error*. As such, their measure of predictability is defined with respect to a forecast system. If the forecast system is *perfect*, by which we mean that the forecast error is distributed as $p(x_{t+\tau}|o_t)$, aside from a shift in mean, then P_o defined above is formally equivalent to predictive information as defined by Schneider and Griffies. P_o defined above avoids the concept of forecast error by invoking instead the prior and posterior distributions. To avoid confusion, P_o will be called the *predictive information for a perfect model*. If the perfect model scenario is clear from context, we call P_o simply the predictive information.

The first term on the right of (4) measures the uncertainty when no observation about the system is available, while the second term measures the uncertainty after the observation becomes available. The difference in uncertainties quantifies the amount of information about the event gained from observation. Schneider and Griffies (1999) review the properties of predictive information, to which we refer the reader for full details. Here we briefly state two properties. First, in the continuous limit, predictive information is invariant with respect to a nonsingular, linear coordinate transformation. This property means that predictive information is independent of the basis set used to represent the data. It follows that all climate variables, regardless of their units or natural variances, can be accounted for in a single measure of predictability. Second, for Gaussian variables, predictive information can be decomposed easily into a set of uncorrelated components ordered by their entropy difference, such that the leading component is the most predictable, the second is the most predictable uncorrelated with

the first, and so on. This decomposition is explained in section 5. Decomposing a metric such as mean square error into components that optimize error can fail to detect highly predictable components if the system is dominated by a background random noise field.

Predictive information is not always positive– the entropy of a conditional distribution can exceed that of the unconditional distribution. Table 1 demonstrates this fact by example. This situation reflects the fact that, sometimes, a small amount of data leads to "confusion." However, the *average* predictive information is positive. To show this, we need to distinguish between the entropy of a conditional distribution for a given value of o_t, and the average of this quantity. These quantities are defined as follows:

$$H(X \mid O = o_t) = -\sum_x p(x \mid o_t) \log\left(p(x \mid o_t) \right)$$
$$H(X \mid O) = \sum_{o_t} p(o_t)\, H(X \mid O = o_t)\ . \qquad \textbf{(5)}$$

The former quantity depends on the realization o_t, whereas the latter quantity, which averages over o_t, does not. The latter quantity is called the *conditional entropy* and satisfies the inequality $H(X) \geq H(X|O)$ (Cover and Thomas 1991, ch. 2). This inequality implies that the *average* predictive information is positive– observations add information on average. For *normal* distributions, predictive information is positive (appendix 1).

An alternative measure of the difference between the posterior and prior distributions is the difference in their information:

$$h(x_{t+\tau}) - h(x_{t+\tau} \mid o_t) = \log\left(\frac{p(x_{t+\tau} \mid o_t)}{p(x_{t+\tau})} \right). \qquad \textbf{(6)}$$

This difference depends on the realization $x_{t+\tau}$, of which there could be many. A summary measure is provided by the average information difference, weighted by the probability of event $x_{t+\tau}$. Of the two possibilities for this weighting, the most useful is:

$$R(p(x_{t+\tau} \mid o_t) \,\|\, p(x_{t+\tau})) = \sum_{x_{t+\tau}} p(x_{t+\tau} \mid o_t) \log\left(\frac{p(x_{t+\tau} \mid o_t)}{p(x_{t+\tau})} \right). \qquad \textbf{(7)}$$

The above quantity, called *relative entropy*, was proposed by Kleeman (2002) as a measure of predictability. Relative entropy also is known as the Kullback-Leibler distance. Kullback (1959) gives compelling arguments that the relative entropy is a measure of the difficulty of discriminating between two distributions. Remarkably, relative entropy has all of the properties listed above for predictive information– it is positive, invariant with respect to linear transformation (in the continuous limit), and decomposable into components that optimize it (as discussed in sec. 5).

Kleeman (2002) points out that the relative entropy of a stationary, Markov process varies monotonically, whereas absolute entropy does not. This statement deserves comment. First, Cover and Thomas (1991, ch. 2) show that the *conditional entropy* $H(X_{t+\tau} \mid X_t)$ of a stationary, Markov process increases. It follows that the predictive information of a stationary, Markov process decreases monotonically *on*

average. Second, if the distributions are *normal*, then the entropy of a stationary, Markov process always increases. This fact is proven in appendix 2. These facts suggest that increases in predictive information are possible but not frequent. Third, Cover and Thomas (1991, ch. 2), to which Kleeman refers, prove that the relative entropy between a stationary, Markov process and *any* stationary distribution decreases monotonically with time. Since the stationary reference can be arbitrary, decreases in relative entropy cannot be interpreted literally to mean that the "distance" between two distributions decreases. In fact, relative entropy does not satisfy the triangle inequality, so a decrease in relative entropy does not necessarily imply that two distributions are getting "closer."

Predictive information and relative entropy have different interpretations in the context of coding theory. For discrete distributions, entropy is the average number of bits (i.e., the average number of "yes-no" questions) needed to describe a variable from a distribution. Hence the difference in entropy is the average reduction in number of bits between a code derived from the prior, versus a code derived from the posterior. In order to describe variables from the prior distribution, a code derived from the prior would have average length $H(x_{t+\tau})$, whereas a code derived from the posterior would have average length $H(x_{t+\tau}) + R(p(x_{t+\tau}|o_t)||p(x_{t+\tau}))$. Thus, relative entropy measures the code efficiency; it is the average increase in the number of bits used to describe a state drawn from the prior distribution, based on a code developed from the posterior.

A key difference between the above two measures is that relative entropy (7) vanishes if *and only if* the two distributions are equal, whereas this is not the case for

predictive information (4). In climate change problems, there are few constraints to prevent predictive information from vanishing even when the distributions differ, whereas relative entropy never vanishes if the distributions differ.

The above measures depend on the observation used for conditioning. One can conceive of another measure of predictability derived by averaging over all observations. It can be shown that both relative entropy and predictive information have the same such average. To see this explicitly, the average relative entropy can be computed as follows:

$$\begin{aligned} \sum_o p(o)\ R(\,p(x\,|\,o)\,\|\,p(x)\,) &= \sum_o p(o) \sum_x p(x\,|\,o)\ \log\left(\frac{p(x\,|\,o)}{p(x)}\right) \\ &= \sum_o \sum_x p(x\,,\,o)\,\log\left(\frac{p(x\,,\,o)}{p(x)\,p(o)}\right), \end{aligned} \tag{8}$$

where the subscripts indicating time have been suppressed, $p(x\,,\,o)$ is the joint distribution between x and o, and the classical relation $p(x\,,\,o) = p(x\,|\,o)\,p(o)$ has been used. Similarly, the average predictive information is

$$\begin{aligned} \sum_o p(o)\,P_O &= -\sum_x p(x)\,\log p(x) + \sum_o p(o) \sum_x p(x\,|\,o)\,\log p(x\,|\,o) \\ &= -\sum_{o\,,\,x} p(x\,,\,o)\,\log p(x) + \sum_{o\,,\,x} p(x\,,\,o)\,\log\left(\frac{p(x\,,\,o)}{p(o)}\right) \\ &= \sum_{o\,,\,x} p(x\,,\,o)\,\log\left(\frac{p(x\,,\,o)}{p(o)\,p(x)}\right), \end{aligned} \tag{9}$$

where we have used the classical relation

$$p(x) = \sum_{o} p(x, o) . \qquad \textbf{(10)}$$

Comparison between (8) and (9) show that the quantities have the same average. Interestingly, not only do relative entropy and predictive information have the same average, but this average is precisely the measure of predictability introduced by Leung and North (1990) called transinformation. Cover and Thomas (1991) call this quantity mutual information, a terminology we adopt in this paper. Mutual information measures the dependence between two variables and is defined as

$$I(x ; o) = \sum_{o, x} p(x, o) \log\left(\frac{p(x, o)}{p(o)\, p(x)} \right) . \qquad \textbf{(11)}$$

which can be seen to equal (8) and (9). Cover and Thomas (1991, ch. 2) show that mutual information $I(x_{t+\tau}, x_t)$ decreases monotonically for a stationary, Markov process.

The problem of comparing two distributions arises not only in predictability, but also in statistical hypothesis testing. It is therefore of interest to compare how the approach based on information theory differs from that based on statistics. A routine example is that of deciding which of several possible distributions a given sample was drawn. The optimum test, in the sense of having desirable properties with respect to type I and type II errors, is given by the Neyman-Pearson liklihood ratio test, which is

formally equivalent to comparing the relative entropy between the empirical distribution and the candidate distributions (Kullback 1959). As another example, Pearson's chi-square test is a classical method of deciding whether two distributions differ. It can be shown that the chi-square statistic is formally equivalent to relative entropy in the limit of small differences between the two distributions. An important problem in time series analysis is determining the order of an autoregressive model which is most appropriate for simulating an observed time series. One of the most used criteria is known as the Akaike Information Criteria (AIC, Brockwell and Davis 1991, von Storch and Zwiers 1998). AIC is derived by determining the expected value of the relative entropy between the model distribution and the "truth," where the truth is unknown but irrelevant since it does not affect differences in AIC. Another basis for quantifying the difference between two distributions is by the amount of money a gambler would make after observations become available, relative to a "bookie" who assigned fair odds prior to observations. In effect, Cover and Thomas (1991, sec. 6.1) show that average doubling rate of the gamblers winnings equals the relative entropy between the posterior and prior distributions. These examples illustrate that information theory often suggests statistical measures that are potentially of financial interest, and are not far removed from those suggested by statistical theory.

4 Predictability of Stochastic Models

Stochastic models provide an opportunity to understand the predictability of a system comprehensively. Moreover, Wang and Uhlenbeck (1945) show that stochastic models mimic *any* stationary, Gaussian, Markov process. That is, a stochastic model provides an equivalent description of such processes in physically meaningful terms; i.e., in terms of a dynamical system driven by noise. A linear stochastic model is of the form

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{w}\,, \tag{12}$$

where $\boldsymbol{x}$ is a state vector, the dot indicates a time derivative, $\boldsymbol{A}$ is a dynamical operator, and $\boldsymbol{w}$ is a Gaussian white noise process with zero mean and covariance matrix $\boldsymbol{Q}$. If the dynamical operator is independent of time, then the system is autonomous and the solution can be calculated analytically. Such models are well understood and have been reviewed in numerous publications (Gardiner 1990; DelSole 2003). Here we simply state relevant results. The initial condition $\boldsymbol{x}_t$ and the "verification" $\boldsymbol{x}_{t+\tau}$ represent solutions of (12) for a single realization of the noise. If the stochastic process was begun in the infinite past, then $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+\tau}$ are stationary processes with distribution

$$p(\boldsymbol{x}_t) = p(\boldsymbol{x}_{t+\tau}) = N(\boldsymbol{0}\,,\,\boldsymbol{\Sigma}_v)\,, \tag{13}$$

where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and the "climatological" covariance matrix is given by

$$\boldsymbol{\Sigma}_v = \int_0^{\tau} e^{\boldsymbol{A}s}\, \boldsymbol{Q}\, e^{\boldsymbol{A}^T s}\, ds\,. \qquad \textbf{(14)}$$

The distribution $p(\boldsymbol{x}_{t+\tau})$ can be interpreted as the prior distribution when no (recent) observation is available. The initial condition $\boldsymbol{x}_t$ and verification $\boldsymbol{x}_{t+\tau}$ are related by

$$\boldsymbol{x}_{t+\tau} = e^{\boldsymbol{A}\tau}\boldsymbol{x}_t + \boldsymbol{\eta}_v(t+\tau) \qquad \textbf{(15)}$$

where $\boldsymbol{\eta}_v$ is distributed as

$$\boldsymbol{\eta}_v(t+\tau) \sim N\left(\mathbf{0}\,, \boldsymbol{\Sigma}_v - e^{\boldsymbol{A}\tau}\boldsymbol{\Sigma}_v e^{\boldsymbol{A}^T\tau}\right) \qquad \textbf{(16)}$$

The first term on the right of (15) constitutes an unbiased estimate of $\boldsymbol{x}_{t+\tau}$, and is not a random variable when $\boldsymbol{x}_t$ is known. Schneider and Griffies (1999) write a "constitutive equation" of the same form as (15) (the first displayed equation in their paper) and call the first term on the right the predictor and the second term the error. Kleeman (2002) in essence calls the first term the signal and the second term the noise.

The conditional distribution at time $t + \tau$, given knowledge of the state at time t, is

$$p(\boldsymbol{x}_{t+\tau} \mid \boldsymbol{x}_t) \sim N\left(e^{\boldsymbol{A}\tau}\boldsymbol{x}_t\,, \boldsymbol{\Sigma}_v - e^{\boldsymbol{A}\tau}\boldsymbol{\Sigma}_v e^{\boldsymbol{A}^T\tau}\right). \qquad \textbf{(17)}$$

This distribution can be interpreted as the posterior distribution when the initial condition becomes known. Importantly, the initial condition $\boldsymbol{x}_t$ affects only the mean of the posterior. In the limit of large lead time, the posterior (17) approaches the prior (13).

The relative entropy for normal distributions is given in Kullback (1959; ch. 9). Substituting into that expression the above prior and posterior distributions, and invoking simple properties of the trace and determinant, gives:

$$R\Big(p(\boldsymbol{x}_{t+\tau}\,|\,\boldsymbol{x}_t)\,\|\,p(\boldsymbol{x}_{t+\tau})\Big) = -\frac{1}{2}Tr[\,\boldsymbol{W}\,\boldsymbol{W}^{T}\,] - \frac{1}{2}\log|\boldsymbol{I} - \boldsymbol{W}\,\boldsymbol{W}^{T}| + \frac{1}{2}\boldsymbol{x}_t^{T}\Big(\boldsymbol{\Sigma}_v^{-1/2}\,\boldsymbol{W}^{T}\,\boldsymbol{W}\,\boldsymbol{\Sigma}_v^{-1/2}\Big)\boldsymbol{x}_t\,, \tag{18}$$

where

$$\boldsymbol{W} = \boldsymbol{\Sigma}_v^{-1/2}\,e^{\boldsymbol{A}\tau}\,\boldsymbol{\Sigma}_v^{1/2}\,. \tag{19}$$

The matrix $\boldsymbol{W}$ is called the pre-whitened dynamical operator. Kleeman (2002) calls the last term in (18) the "signal." The initial condition $\boldsymbol{x}_t$ affects only the last term in (18).

The entropy for multivariate normal distributions is given in Cover and Thomas (1991, eq. 9.45) and Schneider and Griffies (1999, pg. 3137). Substituting into that expression the above prior and posterior distributions gives the predictive information

$$H(\boldsymbol{X}_{t,\tau}) - H(\boldsymbol{X}_{t,\tau} \mid \boldsymbol{X}_t = x_t) = -\frac{1}{2}\log\left(|\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T|\right). \qquad \textbf{(20)}$$

Note that the above expression is independent of the initial condition, in contrast to the relative entropy (18). Finally, the mutual information between the verification and initial condition can be obtained by averaging the predictive information (20) over all initial conditions, which is a trivial exercise with the result

$$I(\boldsymbol{x}_{t,\tau}; \boldsymbol{x}_t) = -\frac{1}{2}\log\left(|\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T|\right). \qquad \textbf{(21)}$$

A theorem in Tippett and Chang (2003) shows that, for all three measures of predictability above, the least predictable system, out of all systems with the same eigenvalues and vanishing initial error, is that in which the pre-whitened dynamical operator $\boldsymbol{W}$ is *normal*. Since the pre-whitened operator is unique up to an orthogonal transformation, the normality of $\boldsymbol{W}$ is not altered by admissible coordinate changes.

To connect with classical theory, we consider a measure of predictability based on the mean square difference between two randomly chosen members of the posterior distribution. If $\boldsymbol{\varepsilon}$ is the difference between two randomly chosen states, and $\boldsymbol{M}$ is a positive definite matrix defining the norm, then it can be shown that

$$\begin{aligned} \{\boldsymbol{\varepsilon}^T \boldsymbol{M} \boldsymbol{\varepsilon}\} &= 2\ Tr\left\{\boldsymbol{M}\left(\boldsymbol{\Sigma}_v - e^{\boldsymbol{A}\tau}\boldsymbol{\Sigma}_v e^{\boldsymbol{A}^T\tau}\right)\right\} \\ &= 2\ Tr\left\{\left(\boldsymbol{\Sigma}_v^{1/2}\boldsymbol{M}\boldsymbol{\Sigma}_v^{1/2}\right)\left(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^T\right)\right\}. \end{aligned} \qquad \textbf{(22)}$$

where the brackets denote an average over the posterior distribution (over the "forecast ensemble"). Note that this measure is independent of the initial condition.

Interestingly, the lead-time dependence of all the above measures are controlled by the operator $\boldsymbol{W}$. Appendix 2 shows that all of the above measures are monotonic functions of lead time, and hence provide interpretations of the second law of thermodynamics for normally distributed, stationary, Markov processes.

By far the most significant difference among the above measures is that predictive information, mutual information, and mean square "error" are independent of the initial condition, whereas the relative entropy is not. A revealing thought experiment is to consider the case in which the initial condition projects on the leading singular vector of a suitable propagator. In this case, the ensemble mean grows to a large value. Despite this mean growth, predictive information gives precisely the same value regardless of the state of the initial condition. This thought experiment may not be academic– Penland and Sardeshmukh (1995) and Moore and Kleeman (1999) have recently proposed that ENSO events can be understood in precisely this way. The mere possibility of such a theory, regardless of its correctness, reveals that predictive information and mean square error fail to capture a fundamental component of predictability.

5 Predictable Components

Reducing all information about predictability to a single number, say relative entropy, is a gross simplification. To gain further insight, relative entropy may be decomposed into a set of components that optimize this quantity. This approach is analogous to the use of principal component analysis to understand how different spatial structures contribute to total variance. Decomposition of "error variance" would not be as useful because error variance may be "small" not only because the error is small, but because the associated spatial structures have small natural variances. Schneider and Griffies (1999) show that, in the case of Gaussian variables, the decomposition of predictive information may be reduced to an eigenvalue problem. In this section, we apply a similar decomposition technique to relative entropy and mutual information. It will be shown that the decomposition of predictive information is identical to the decomposition of relative entropy, in the case in which the means of the two distributions are identical. The optimization of predictive information, called predictable component analysis, is discussed in Schneider and Griffies (1999), to which the reader is referred for details (see also DelSole and Chang 2003). Here we give only an outline of the method.

It proves convenient to distinguish the prior and posterior distributions by different variables. Thus, the variable having the prior distribution, called the verification, will be denoted by $\boldsymbol{x}_v$, while the variable with the posterior distribution will be denoted by $\boldsymbol{x}_{v|i}$. To understand how individual components contribute to predictive information, we determine a projection vector $\boldsymbol{q}_k$ such that the projected quantities

optimize predictive information. The subscript k is an index anticipating the fact that multiple vectors will be obtained. Denote the projected variables as $\varepsilon_{v,k}(t)=\boldsymbol{q_k}^T\boldsymbol{x_v}$ and $\varepsilon_{f,k}(\tau,t)=\boldsymbol{q_k}^T\boldsymbol{x_{v|i}}$, both of which are scalar time series. If the variables are Gaussian, then these scalar time series also follow a Gaussian distribution, but with *scalar* means $\boldsymbol{q_k}^T\boldsymbol{\mu_v}$ and $\boldsymbol{q_k}^T\boldsymbol{\mu_{v|i}}$ and variances $\boldsymbol{q_k}^T\boldsymbol{\Sigma_v}\boldsymbol{q_k}$ and $\boldsymbol{q_k}^T\boldsymbol{\Sigma_{v|i}}\boldsymbol{q_k}$. These scalars will be denoted as

$$\begin{aligned} \mu_{v|i} &= \boldsymbol{q_k}^T\boldsymbol{\mu_{v|i}} \qquad \sigma^2_{v|i} = \boldsymbol{q_k}^T\boldsymbol{\Sigma_{v|i}}\boldsymbol{q_k} \\ \mu_{v} &= \boldsymbol{q_k}^T\boldsymbol{\mu_{v}} \qquad \sigma^2_{v} = \boldsymbol{q_k}^T\boldsymbol{\Sigma_{v}}\boldsymbol{q_k}, \end{aligned} \tag{23}$$

where the k-dependence of these scalars is understood. The predictive information associated with projection vector $\boldsymbol{q_k}$ can be written as

$$P_k = \frac{1}{2}\log\left(r_k\right) \tag{24}$$

where r_k is the ratio of variances associated with the k'th projection vector

$$r_k = \frac{\boldsymbol{q_k}^T\boldsymbol{\Sigma_{v|i}}\boldsymbol{q_k}}{\boldsymbol{q_k}^T\boldsymbol{\Sigma_{v}}\boldsymbol{q_k}} = \frac{\langle\varepsilon^2_{f,k}\rangle}{\langle\varepsilon^2_{v,k}\rangle} = \frac{\sigma^2_{v|i}}{\sigma^2_v}. \tag{25}$$

Since the log-function is monotonic, optimization of predictive information P_k is equivalent to optimization of the variance ratio r_k. The variance ratio (25) is a Rayleigh

quotient. It is a standard procedure (Noble and Daniel 1988) to show that optimization of the quotient r_k with respect to $\boldsymbol{q_k}$ leads to the generalized eigenvalue problem

$$\boldsymbol{\Sigma}_{v|i}\, \boldsymbol{q_k} = \lambda\, \boldsymbol{\Sigma}_{v}\, \boldsymbol{q_k}\,. \qquad \textbf{(26)}$$

The eigenvectors yield statistically uncorrelated time series; that is, $\{ \varepsilon_{f,k}\, \varepsilon_{f,m} \} = 0$ and $\{ \varepsilon_{v,k}\, \varepsilon_{v,m} \} = 0$ for $k \neq m$. The corresponding eigenvalues give the variance ratio r_k associated with each vector. Ordered by their eigenvalues from largest to smallest, the leading eigenvector gives the projection vector that maximizes predictive information, the second maximizes predictive information over all vectors uncorrelated with the first, and so on.

Now consider the determination of components that optimize relative entropy. The relative entropy R_k associated with projection vector $\boldsymbol{q_k}$ is

$$R_k = \frac{1}{2}\left(\log\left(\frac{\sigma_v^2}{\sigma_{v|i}^2} \right) + \frac{\sigma_{v|i}^2}{\sigma_v^2} - 1 + \frac{(\mu_{v|i} - \mu_v)^2}{\sigma_v^2} \right), \qquad \textbf{(27)}$$

where the terms appearing on the right are defined in (23). All terms appearing in (27) are ratios of quadratic forms. It follows that R_k is invariant with respect to multiplicative factors applied to the projection vector $\boldsymbol{q_k}$. Without loss of generality, the projection vectors may be normalized such that $\sigma_v^2 = 1$, where σ_v^2 is given in (23). Optimization of R_k with respect to $\boldsymbol{q_k}$, subject to the constraint $\sigma_v^2 = 1$, leads to the equation

$$\left(\boldsymbol{\Sigma}_{v|i} + \frac{\left(\boldsymbol{\mu}_{v|i} - \boldsymbol{\mu}_v\right)\left(\boldsymbol{\mu}_{v|i} - \boldsymbol{\mu}_v\right)^T}{1 - 1/\sigma^2_{v|i}} \right) \boldsymbol{q}_k = \lambda \boldsymbol{\Sigma}_v \boldsymbol{q}_k , \qquad \textbf{(28)}$$

where λ is a Lagrange multiplier. Since $\sigma_{v|i}{}^2$ depends on $\boldsymbol{q}_k$ (see (23)), the above equation is nonlinear and solvable only by iterative procedures. Nevertheless, it still holds that if $\boldsymbol{q}_k$ and $\boldsymbol{q}_j$ are two solutions of (28) with different values of λ, then $\boldsymbol{q}_k{}^T \boldsymbol{\Sigma}_v \boldsymbol{q}_j = \delta_{k\,j}$; that is, the time series associated with $\boldsymbol{q}_k$ and $\boldsymbol{q}_j$ are uncorrelated with respect to the verification ensemble. This may be shown in essentially the same way that one proves that eigenvectors of symmetric matrices with distinct eigenvalues are orthogonal.

We now come to an interesting fact: if $\boldsymbol{x}_v$ and $\boldsymbol{x}_{v|i}$ have the same means, then (28) reduces to the eigenvalue problem (26), from which it follows that relative entropy and predictive information share the same predictable components. This correspondence is not surprising for two normal distributions with identical means, since such distributions differ only in their variances. In the opposite extreme in which covariances are nearly the same, the matrix on the left of (28) is dominated by the terms involving the means, in which case the projection vector can be solved analytically as

$$\boldsymbol{q}_k \propto \boldsymbol{\Sigma}_v^{-1}\left(\boldsymbol{\mu}_{v\,|\,i} - \boldsymbol{\mu}_v\right). \qquad \textbf{(29)}$$

This projection vector also arises as the “fingerprint” for optimizing the signal to noise ratio of the difference in means (Hasselmann 1993).

Now consider the determination of components that optimize mutual information. The mutual information for multivariate normal distributions is

$$I(\boldsymbol{x}_v ; \boldsymbol{x}_o) = -\frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_v - \boldsymbol{\Sigma}_{vo}\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\Sigma}_{ov}|}{|\boldsymbol{\Sigma}_v|}\right), \qquad \textbf{(30)}$$

where $\boldsymbol{\Sigma}_{ov} = \boldsymbol{\Sigma}_{vo}^{\mathrm{T}}$ is the cross-covariance matrix between the observation and verification, $\boldsymbol{x}_o$ and $\boldsymbol{x}_v$. We follow the same procedure as above, except that we introduce a *pair* of projection vectors, say $\boldsymbol{q}_k$ and $\boldsymbol{s}_k$. Let the projected initial condition be $\varepsilon_{i\,k}(\tau,t)=\boldsymbol{q}_k^T\boldsymbol{x}_o$, and the projected verification be $\beta_{v,k}(t)=\boldsymbol{s}_k^T\boldsymbol{x}_v$, both of which are scalar time series. The projection renders all covariance matrices in (30) scalar variables, which can be manipulated to show that the mutual information between $\varepsilon_{i\,k}$ and $\beta_{v,k}$ is

$$I_k = -\frac{1}{2}\log\left(1-\rho_k^2\right), \qquad \textbf{(31)}$$

where ρ_k is the cross-correlation between the two scalar variables. Thus, for Gaussian variables, the projection vector pair that maximizes I_k also maximizes the squared correlation between the associated time series. This relation is reasonable in view of the fact that mutual information measures the degree of dependence between two variables, which in the case of Gaussian variables is quantified by ρ. However, optimization of correlation is solved by a well known procedure called Canonical Correlation Analysis

(CCA). It follows that, in the case of Gaussian variables, CCA applied to the initial and verification fields determines the components that maximize mutual information.

The power of the above decompositions become evident when predictability is dominated by a few coherent patterns. In such cases, the decomposition allows an investigator to focus only on the few structures which are predictable and to ignore the vast majority of structures which are unpredictable. This decomposition also provides the basis for attributing predictability to specific structures in the initial condition, boundary condition, and/or external forcing. This is called the *attribution* problem. The basic idea is to compute the conditional probability of the verification given partial knowledge of the state of the system. For instance, one could consider the conditional distribution given the initial condition only over certain geographic regions, or given the sea surface temperature only over certain geographic regions. Predictable component analysis applied to such scenarios would then yield specific structures in the verification whose predictability can be attributed directly to the knowledge used for conditioning.

Decomposition of mutual information offers a systematic approach to the attribution problem since it yields a complete set of observation-verification pairs which dominate the *average* predictability, thereby avoiding the need to explore each observation individually. Furthermore, it is possible to produce forecasts based only on the predictable components (see Barnett and Preisendorfer (1987) and DelSole and Chang (2003) for a review of the CCA procedure and the method for constructing predictions based on canonical patterns).

6 Summary and Discussion

This paper reviewed a framework for understanding predictability based on information theory. The fundamental quantity in predictability theory is the conditional distribution, which gives the distribution of the state conditional on knowledge about the system gained by observation. An event is said to be predictable if the distribution prior to the observation differs in any way from the distribution after the observation becomes available. The degree of predictability of an event therefore depends on the degree to which the prior and posterior distributions differ. Information theory provides an intuitive approach to measuring the difference between two distributions, because it ascribes more predictability to events which have less uncertainty in their prediction.

There are, however, different ways in which to measure the difference in uncertainty. Schneider and Griffies (1999) propose measuring predictability by the predictive information– that is, the entropy difference between the prior and posterior distributions (for a perfect forecast model). Kleeman (2002) proposed measuring predictability by relative entropy– the average difference in information between the prior and posterior distributions. Finally, Leung and North (1990) proposed measuring predictability by mutual information-- the degree of dependence between the verification and observation. All three quantities have several attractive properties in the context of predictability theory. First, all three quantities are invariant with respect to a nonsingular linear transformation. This invariance means that the metric does not depend on the basis in which the data is represented, and also that variables with different units and

different natural variances can be analyzed as a single state variable without introducing an arbitrary normalization, since such normalization cannot affect the final results. This property is not shared by such metrics as mean square error. Second, for Gaussian variables, all three quantities can be decomposed easily into a set of uncorrelated components which optimize these quantities. The components can be ordered by their measure of predictability such that the leading component maximizes the predictability over all possible components, the second maximizes predictability over all components uncorrelated with the first, and so on. Third, if the underlying system is stationary and Markovian, then all three quantities are a non-increasing function of lead time *in some sense*. Predictive information has this property only in an average sense. The fact that predictive information does not satisfy this property in all cases reflects the fact that, sometimes, a small set of observations can lead to "confusion." However, this confusion is rare, in the sense that, *on average*, data reduces the entropy. Relative entropy decreases monotonically without averaging, but the interpretation of this property is obscure owing to the fact that relative entropy is not a true distance, since it depends on the order of the distributions and does not satisfy the triangle inequality. The appropriate monotonic decay of both predictive information and relative entropy provide interpretations of the second law of thermodynamics and conforms with the classical notion (Lorenz 1969) that predictability degrades with time.

The fact that all three metrics share numerous properties raises the question as to which metric should be used to define predictability. Relative entropy and predictive information both measure the predictability of a *single* prediction. Mutual information

should be sharply distinguished from the other two metrics in that it measures the *average* of these two quantities over all predictions. Hence, mutual information can be interpreted as a measure of the predictability expected from a "typical" observation. The most fundamental difference between relative entropy and predictive information is that relative entropy vanishes if and only if the two distributions are identical, whereas this is not the case for predictive information. This fact appears to confer an advantage to relative entropy. As Kleeman (2002) emphasized in the context of normal distributions, relative entropy depends on the difference in means whereas predictive information does not. A revealing thought experiment is to consider the predictability of a linear stochastic model. In this case, the entropy of the system increases in time due to the spread of the ensemble, but is completely independent of the initial condition. If the initial condition projects on the leading singular vector of the propagator, both the mean and spread of the ensemble grow, but the predictive information gives the same value regardless of the ensemble mean, in contrast to relative entropy. This thought experiment gives a clear example in which relative entropy captures a potentially important component of predictability that would not be captured by predictive information, or by the classical predictability measure based on the mean square forecast error.

All three quantities can be decomposed into components that optimize these quantities. This decomposition is very useful in the case in which predictability is dominated by a few components, since it allows an investigator to focus only on the few structures which are predictable. These decompositions also provide the basis for attributing predictability to specific structures in the initial condition, boundary

condition, and/or external forcing through controlled experiments. Mutual information offers an attractive approach to the attribution problem, since its decomposition yields a complete set of observation-verification pairs which dominate the average predictability, thereby avoiding the need to explore each initial condition individually. Obviously, the decomposition can be applied not only to the initial condition, but also to boundary conditions and other knowledge that may be available. Furthermore, it is possible to identify the predictable components and then produce statistical forecasts based only on these components. If the variables are Gaussian, then the decomposition of mutual information is equivalent to canonical correlation analysis applied to the observation and verification; the canonical patterns can be interpreted as components that optimize the average predictability. Finally, in the case in which there is no change in "internal noise," optimization of relative entropy leads to optimal detection and "fingerprint" methods previously employed in climate change studies. In this sense, the above methods can be interpreted as a more general framework for quantifying climate change since they account not only for changes in the mean, but also changes in the variability.

A fundamental limitation of the above methodology is that the required probability distributions are not known and must be estimated from data. In practice, this problem is dealt with by using forecast models to simulate the posterior distribution. Unfortunately, errors in all realistic forecast models give rise to differences between the forecast and truth which, if not accounted for, would lead to a false conclusion about the existence of predictability. The question as to how to account for forecast errors in the context of information theory will be discussed in part II of this paper.

Acknowledgments

I am very much indebted to J. Shukla, who provided a perfect balance between criticism and encouragement during the course of this work. I also thank T. Schneider, M. Tippett and B. Kirtman for instructive discussions. This research was supported by the NSF (ATM9814295), NOAA (NA96-GP0056), and NASA (NAG5-8202).

7 Appendix I: The Sign of Predictive Information

In this appendix we show that the predictive information between unconditional and conditional distributions is positive if the variables are normally distributed. Let the P-dimensional vector $\boldsymbol{X}$ be the variables of interest, and let $\boldsymbol{Y}$ be the observed variables; these two vectors do not necessarily have the same dimension nor represent the same variables. Let the mean and covariance matrix of $\boldsymbol{X}$ be $\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_{xx}$, and those of $\boldsymbol{Y}$ be $\boldsymbol{\mu}_y$ and $\boldsymbol{\Sigma}_{yy}$, respectively, and let the cross-covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$ be $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}^{\mathrm{T}}$. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ follow a joint normal distribution, then the conditional distribution, given that the random variable $\boldsymbol{Y}$ equals $\boldsymbol{y}$, is (Johnson and Wichern 1982)

$$p\left(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}\right) = N_p\left(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_y\right),\ \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right), \qquad \textbf{(32)}$$

where $N_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined in (13). The entropy of a normal distribution is given explicitly in Cover and Thomas (1991, ch. 9) and Schneider and Griffies (1999). Using this expression, the entropy of the conditional distribution (32) is

$$H(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{2}\left(P \log 2\pi + P + \log \mid \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx} \mid \right) \tag{33}$$

and the entropy of the unconditional distribution is

$$H(\boldsymbol{X}) = \frac{1}{2}\left(P \log 2\pi + P + \log \mid \boldsymbol{\Sigma}_{xx} \mid \right). \tag{34}$$

It follows that the predictive information is

$$\begin{aligned} P_y &= H(\boldsymbol{X}) - H(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{2} \log \left(\frac{\mid \boldsymbol{\Sigma}_{xx} \mid}{\mid \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx} \mid} \right) \\ &= -\frac{1}{2} \log \left(\mid \boldsymbol{I} - \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx} \mid \right), \end{aligned} \tag{35}$$

where standard properties of the determinant have been used to derive the last expression. The matrix inside the determinant has well known properties owing to its fundamental role in canonical correlation analysis. In particular, the eigenvalues of $\boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx}$ lie between 0 and 1. Since the determinant of a matrix equals the product of eigenvalues, it follows that the predictive information R_y is non-negative. Thus, we have shown that the predictive information of a normally distributed set of variables, given an observation that reveals the exact value of a set of variables, is non-negative. The above result may be interpreted as proving that *any* conditioning of a normal distribution results in a decrease in entropy.

8 Appendix II: The Variation of Predictability

Cover and Thomas (1991, ch. 2) show that mutual information and relative entropy decay monotonically for discrete, stationary, Markov chains. This result is extended here to show that, for continuous normal distributions of stationary, Markov processes, relative entropy, predictive information, mutual information, and mean square error are monotonic functions of lead time.

A basic result which will be needed is the derivative of the logarithm of a determinant. It can be shown that for any non-singular matrix $\boldsymbol{X}$,

$$\frac{d}{dt}\log|\boldsymbol{X}| = Tr\left\{\boldsymbol{X}^{-1}\frac{d\boldsymbol{X}}{dt}\right\}. \tag{36}$$

This result can be derived by considering $\log|\boldsymbol{X}|$ at two infinitesimally close values of t and invoking standard properties of determinants and log-functions in the differential. It will also prove convenient to define the following quantities

$$\begin{aligned}\boldsymbol{\Phi} &= \boldsymbol{I} - \boldsymbol{\Sigma}_v^{-1} e^{\boldsymbol{A}\tau}\boldsymbol{\Sigma}_v e^{\boldsymbol{A}^T\tau}\\ \boldsymbol{\Gamma} &= \boldsymbol{\Sigma}_v^{-1} e^{\boldsymbol{A}\tau}\boldsymbol{Q}\, e^{\boldsymbol{A}^T\tau}.\end{aligned} \tag{37}$$

It can be shown that both matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ are positive definite for finite τ, and that the eigenvalues of $\boldsymbol{\Phi}$ lie between 0 and 1.

The properties of normally distributed, stationary Markov processes were reviewed in sec. 4, within the equivalent context of linear stochastic models. The derivative of mutual information (21) with respect to lead time τ is given by

$$\begin{aligned} \frac{dI}{d\tau} &= -\frac{1}{2}\frac{d}{d\tau}\log\left|\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^{T}\right| = -\frac{1}{2}\frac{d}{d\tau}\log\left|\boldsymbol{\Phi}\right| \\ &= +\frac{1}{2}Tr\left\{\boldsymbol{\Phi}^{-1}\boldsymbol{\Sigma}_{v}^{-1}e^{A\tau}\left(\boldsymbol{A}\boldsymbol{\Sigma}_{v} + \boldsymbol{\Sigma}_{v}\boldsymbol{A}^{T}\right)e^{A^{T}\tau}\right\} \\ &= -\frac{1}{2}Tr\left\{\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}\right\}, \end{aligned} \tag{38}$$

where we have used (36) and the Lyapunov equation $\mathbf{A}\boldsymbol{\Sigma}_{v} + \boldsymbol{\Sigma}_{v}\mathbf{A}^{T} + \mathbf{Q} = \mathbf{0}$ associated with the stochastic model (12). Since the product of two positive definite matrices is positive definite, the last line in (38) is negative. This proves that the mutual information (21) and predictive information (20) for a normally distributed, stationary Markov process decay monotonically with lead-time.

Now consider the relative entropy given in (18). The above result proves that the second term in (19) decays monotonically. The derivative of the first term in (18) is

$$\begin{aligned} -\frac{1}{2}\frac{\partial}{\partial\tau}Tr\left\{\boldsymbol{W}\boldsymbol{W}^{T}\right\} &= -\frac{1}{2}\frac{\partial}{\partial\tau}Tr\left\{\boldsymbol{\Sigma}_{v}^{-1}e^{A\tau}\boldsymbol{\Sigma}_{v}e^{A^{T}\tau}\right\} \\ &= -\frac{1}{2}Tr\left\{\boldsymbol{\Sigma}_{v}^{-1}e^{A\tau}\left(\boldsymbol{A}\boldsymbol{\Sigma}_{v} + \boldsymbol{\Sigma}_{v}\boldsymbol{A}^{T}\right)e^{A^{T}\tau}\right\} \\ &= +\frac{1}{2}Tr\left\{\boldsymbol{\Sigma}_{v}^{-1}e^{A\tau}\boldsymbol{Q}\,e^{A^{T}\tau}\right\} = \frac{1}{2}Tr\left\{\boldsymbol{\Gamma}\right\}. \end{aligned} \tag{39}$$

Note that this term increases with lead time. Despite the increase in this term, the sum of the first two terms in (18) decrease with lead time. This follows from the ordering theorem (Noble and Daniel, ch. 7), which states that $\mathbf{\Gamma} - \mathbf{\Gamma}\,\mathbf{\Phi}$ is positive definite if $\mathbf{\Phi}$ has eigenvalues less than unity (which it does). Using a similar procedure, the last term in (18) can be shown to decrease monotonically with lead time. Thus, the sum of all three terms in (18), and hence the relative entropy, decreases monotonically with lead time.

Finally, let us consider the mean square difference between two randomly chosen members of the posterior distribution. If ε is the difference between two randomly chosen members of the posterior distribution (17), and $\boldsymbol{M}$ is a positive definite matrix defining the norm, then

$$\begin{aligned} \frac{d}{d\tau}\left\{\boldsymbol{\varepsilon}^T \boldsymbol{M}\,\boldsymbol{\varepsilon}\right\} &= 2\,\frac{d}{d\tau}\,Tr\left\{\boldsymbol{M}\left(\boldsymbol{\Sigma}_v - e^{\boldsymbol{A}\tau}\boldsymbol{\Sigma}_v\, e^{\boldsymbol{A}^T\tau}\right)\right\} \\ &= -2\,Tr\left\{\boldsymbol{M}\,e^{\boldsymbol{A}\tau}\left(\boldsymbol{A}\boldsymbol{\Sigma}_v + \boldsymbol{\Sigma}_v\boldsymbol{A}^T\right)e^{\boldsymbol{A}^T\tau}\right\} \\ &= 2\,Tr\left\{\boldsymbol{M}\,e^{\boldsymbol{A}\tau}\,\boldsymbol{Q}\,e^{\boldsymbol{A}^T\tau}\right\}, \end{aligned} \qquad \textbf{(40)}$$

where the brackets denote an average over the posterior distribution. The last line in (40) is the trace of two positive definite matrices, and hence is positive. This demonstrates that the mean square difference between two randomly chosen members of the posterior distribution increases monotonically with lead time, *for any norm*.

References

Barnett, T. P., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev*., **115**, 1825–1850.

Brockwell, P. J., and R. A. Davis, 1991: *Timeseries: Theory and Methods*. 2d ed. Springer-Verlag, 577 pp.

Cover, T. M., and J. A. Thomas, 1991: *Elements of Information Theory*. Wiley, 576 pp.

DelSole T., 2003: Stochastic Models of Quasigeostrophic Turbulence. *Surveys in Geophys*., submitted.

DelSole, T., and P. Chang, 2003: Predictable component analysis, canonical correlation analysis, and autoregressive models. *J. Atmos. Sci*., **60**, 409-416.

Epstein, E. S., 1969: Stochastic dynamic predictions. *Tellus*, **21**, 739–759.

Gardiner, C.W., 1990: *Handbook of Stochastic Methods*. 2d ed. Springer-Verlag, 442pp.

Goldman, S., 1953: *Information Theory*. Prentice Hall, 385pp.

Hasselmann, K., 1993: Optimal finger prints for the detection of time-dependent climate change. *J. Climate*, **6**, 1957-1971.

Jazwinski, A. H., 1970: *Stochastic processes and filtering theory*. Academic press, 376pp.

Johnson, R. A., and D. W. Wichern, 1982: *Applied Multivariate Statistical Analysis*. Prentice-Hall, 594 pp.

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

Kullback, S., 1959: *Information theory and statistics*. Wiley & Sons, 399pp. Republished by Dover 1968.

Leung, L.-Y., and G. R. North, 1990: Information theory and climate prediction. *J. Climate*, **3**, 5-14.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci*., **20**, 130–141.

Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17**, 321-333.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307.

Lorenz, E. N., 1975: Climatic predictability. *The Physical Basis of Climate and Climate Modelling*, B. Bolin et al., Eds., GARP Publication Series, Vol. 16, World Meteorological Organization, 132–136.

Moore, A. M., and R. Kleeman, 1999: Stochastic forcing of ENSO by the intraseasonal oscillation. *J. Climate*, **12**, 1199-1220.

Noble, B., and J. W. Daniel, 1988: *Applied Linear Algebra*. 3d ed. Prentice-Hall, 521pp.

Penland, C., and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *J. Climate*, **8**, 1999-2024.

Reza, F. M., 1961: *An introduction to information theory*. McGraw-Hill.

Schneider, T. and S. M. Griffies, 1999: A conceptual framework for predictability studies. *J. Climate*, **12**, 3133-3155.

Shannon, C. E., 1948: A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 370-423, 623-656.

Tippett, M. K., and P. Chang, 2003: Some theoretical considerations on predictability of linear stochastic dynamics. *Tellus*, **55A**, 148-157

von Storch, H., and F. Zwiers, 1998: *Statistical Analysis in Climate Research*. Cambridge University Press, 528 pp.

Wang, M. C., and G. E. Uhlenbeck, 1945: On the theory of the Brownian motion II. *Rev. Mod. Phys.*, **17**, 323-342.

		x		
	p(x,y)	1	2	p(y)
y	1	1/10	3/10	4/10
	2	2/10	4/10	6/10
	p(x)	3/10	7/10	1

		x	
	p(x|y)	1	2
y	1	1/4	3/4
	2	1/3	2/3

$$H(p(x)) = -\frac{3}{10}\log\left(\frac{3}{10}\right) - \frac{7}{10}\log\left(\frac{7}{10}\right) \approx 0.265$$

$$H(p(x|2)) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) \approx 0.276$$

Table 1: The probabilities of a discrete distribution p(x,y) with negative predictive information. The indices x and y have the values of 1 and 2 only. The top table gives the probabilities of the joint distribution p(x,y) for all possible values of x and y, and those for the marginal distributions p(x) and p(y). The bottom table gives the conditional distribution computed from the definition p(x|y)=p(x,y)/p(y). The entropy of p(x) and p(x|2) are given in the box below the tables and shows that the entropy of the conditional distribution exceeds that of the unconditional distribution.