Predictability and Information Theory. Part II: Imperfect Forecasts

TIMOTHY DELSOLE

George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

(Manuscript received 10 June 2004, in final form 2 February 2005)

ABSTRACT

This paper presents a framework for quantifying predictability based on the behavior of imperfect forecasts. The critical quantity in this framework is not the forecast distribution, as used in many other predictability studies, but the conditional distribution of the state given the forecasts, called the regression forecast distribution. The average predictability of the regression forecast distribution is given by a quantity called the mutual information. Standard inequalities in information theory show that this quantity is bounded above by the average predictability of the true system and by the average predictability of the forecast system. These bounds clarify the role of potential predictability, of which many incorrect statements can be found in the literature. Mutual information has further attractive properties: it is invariant with respect to nonlinear transformations of the data, cannot be improved by manipulating the forecast, and reduces to familiar measures of correlation skill when the forecast and verification are joint normally distributed. The concept of potential predictable components is shown to define a lower-dimensional space that captures the full predictability of the regression forecast without loss of generality. The predictability of stationary, Gaussian, Markov systems is examined in detail. Some simple numerical examples suggest that imperfect forecasts are not always useful for joint normally distributed systems since greater predictability often can be obtained directly from observations. Rather, the usefulness of imperfect forecasts appears to lie in the fact that they can identify potential predictable components and capture nonstationary and/or nonlinear behavior, which are difficult to capture by low-dimensional, empirical models estimated from short historical records.

1. Introduction

DelSole (2004a, hereafter Part I) discussed a framework for quantifying predictability based on information theory. This framework, which is reviewed in the following section, requires probability distributions that are not known and, in practice, are estimated from an imperfect forecast model. The purpose of this paper is to discuss an approach to accounting for imperfect forecasts within the above framework. The basic idea is to use not the forecast itself, but the conditional distribution of the state given the forecast. This idea was suggested by Schneider and Griffies (1999), although our interpretation appears to differ from theirs. The assumptions inherent in this approach are laid out in section 3 and the resulting predictability estimates are shown in section 4 to constitute a lower bound on the true predictability and potential predictability. The mutual information between verification and forecast is argued to be an attractive measure of forecast skill. Section 5 discusses predictable components of imperfect models and their potential significance in practical applications. The above concepts are illustrated in section 6 in the context of stationary, Gaussian, Markov systems. Numerical examples are presented in section 7. Finally, a summary of the results is given in section 8.

This paper considers the ideal case of large samples (large in a sense to be discussed in section 3). Strategies for dealing with small samples will be addressed in Part III.

2. Brief review of predictability

In this section we summarize the predictability framework proposed in Part I. Consider a dynamical system of dimension K. The state of the system at time t is specified by a K-dimensional vector $\mathbf{x}_t$, which specifies a point with coordinates $\mathbf{x}_t$ in the K-dimensional space. If the state is uncertain, it is appropriate to de-

Corresponding author address: Timothy DelSole, Center for Ocean–Land–Atmosphere Studies, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705-3106.
E-mail: delsole@cola.iges.org

scribe the state by the density of possible points in phase space. This density is essentially a probability distribution function and evolves in time in a manner described by Liouville's equation for conservative systems. The distribution of $\mathbf{x}_t$ changes discontinuously after the system is observed. Let the set of all observations up to time t be denoted by $\mathbf{o}_t$. Note that $\mathbf{x}_t$ and $\mathbf{o}_t$ often reside in different spaces. The distribution of the state after observations become available is the conditional distribution $p(\mathbf{x}_t|\mathbf{o}_t)$, whose mean is called the analysis. As is well known from state space estimation theory, the analysis distribution $p(\mathbf{x}_t|\mathbf{o}_t)$ depends on the forecast model and therefore is conditioned on the forecast model.

It proves convenient to distinguish the state at two different times by different symbols. Thus, let the initial condition at time t be $\mathbf{i} = \mathbf{x}_t$, the verification at time $t + \tau$ be $\mathbf{v} = \mathbf{x}_{t+\tau}$, and the observations up to time t be $\mathbf{o} = \mathbf{o}_t$. The parameter τ is called the lead-time. The probability distribution functions (pdf's) of $\mathbf{i}$, $\mathbf{v}$, $\mathbf{o}$ will be denoted by $p(\mathbf{i})$, $p(\mathbf{v})$, $p(\mathbf{o})$, respectively, where the function $p(\cdot)$ is understood to differ according to its argument. In this notation, the analysis distribution at time t is $p(\mathbf{i}|\mathbf{o})$. For stationary systems, $p(\mathbf{v}) = p(\mathbf{i})$.

The distribution of a future state $\mathbf{v} = \mathbf{x}_{t+\tau}$, after observations become available, is denoted by $\mathrm{p}(\mathbf{v}|\mathbf{o})$ and computed from the classical formula:

$$p(\mathbf{v}|\mathbf{o}) = \int r(\mathbf{v}|\mathbf{i})p(\mathbf{i}|\mathbf{o})\, d\mathbf{i}, \tag{1}$$

where $r(\mathbf{v}|\mathbf{i})$ is a *transition probability* associated with a dynamical or stochastic model and the integral is a multiple integral. The distribution $p(\mathbf{v}|\mathbf{o})$ will be called the *perfect model forecast distribution*. This distribution describes our knowledge of the future state $\mathbf{v} = \mathbf{x}_{t+\tau}$ after antecedent observations $\mathbf{o}_t$ and a (perfect model) forecast based on those observations become available. We use the term forecast system to refer to the combined influence of the forecast model and the uncertainty in the initial condition. Note that a perfect model forecast distribution is not "perfectly predictable," even for a deterministic model because even if $r(\mathbf{v}|\mathbf{i})$ is deterministic and hence a delta function, $p(\mathbf{v}|\mathbf{o})$ from (1) is not a delta function, owing to uncertainty in the initial condition as described by $p(\mathbf{i}|\mathbf{o})$.

In the absence of (recent) observations $\mathbf{o}_t$, the variable $\mathbf{v} = \mathbf{x}_{t+\tau}$ has a climatological distribution given by its marginal distribution

$$p(\mathbf{v}) = \int p(\mathbf{v}|\mathbf{o})p(\mathbf{o})\, d\mathbf{o}. \tag{2}$$

If the system is stationary or cyclostationary, then the climatological distribution is independent of time or periodic and can be estimated from historical records. The variable $\mathbf{v}$ is said to be unpredictable if $p(\mathbf{v}|\mathbf{o}) = p(\mathbf{v})$, which is equivalent to the statement that $\mathbf{v}$ is independent of the observations $\mathbf{o}$.

3. The accessible forecast distribution

A key problem with applying the above methodology is that, in practice, the transition probability $r(\mathbf{v}|\mathbf{i})$ for the climate system is not known. It follows then that the perfect model forecast distribution $p(\mathbf{v}|\mathbf{o})$ cannot be computed from (1) and, hence, is unknown too. Moreover, the transition probability $r(\mathbf{v}|\mathbf{i})$ cannot be estimated from data because nature provides only a single realization of $\mathbf{x}_{t+\tau}$ for a given value of $\mathbf{x}_t$, and the atmosphere has no natural analogues in the sense discussed by Lorenz (1969). For these reasons, the transition probability must be estimated from a model. The details of the model are immaterial: for example, the model could be purely empirical or purely physical. What is important is that the model provides a transition probability, which in all realistic cases differs from that of the true system. Moreover, the state space of the forecast model usually differs from that of the true state space. Consequently, the "initial condition" appropriate for the accessible forecast, denoted $\mathbf{i}_f$, differs from the initial condition appropriate for the perfect model forecast, $\mathbf{i}$. Let the initial condition distribution appropriate for the accessible forecast be $p(\mathbf{i}_f|\mathbf{o})$, and let the forecast verifying at time $t + \tau$ be $\mathbf{f}$. The forecast distribution is then given by

$$p(\mathbf{f}|\mathbf{o}) = \int r'(\mathbf{f}|\mathbf{i}_f)p(\mathbf{i}_f|\mathbf{o})\, d\mathbf{i}_f, \tag{3}$$

where $r'(\mathbf{f}|\mathbf{i}_f)$ is the transition probability for the model. The distribution $p(\mathbf{f}|\mathbf{o})$ will be called the accessible forecast distribution, to distinguish it from the perfect model forecast distribution $p(\mathbf{v}|\mathbf{o})$, which is inaccessible in any realistic scenario. Samples drawn from $p(\mathbf{f}|\mathbf{o})$ constitute the forecast ensemble.

Obviously, we would eliminate model errors if we could. Hence, we assume that model errors cannot be eliminated easily. In such situations, there appears to be no alternative other than to quantify predictability based on the past behavior of the model and observations, assuming that the past relation between model and observations will persist into the future. This assumption is reasonable for stationary systems, but is problematic for nonstationary systems, such as occur in climate change scenarios.

Let us assume, then, that the system is stationary. A complete description of the past behavior of the model

and observations is the joint distribution $p(\mathbf{v}, \mathbf{o}, \mathbf{f})$. The distribution of the verification given knowledge of the forecast and observations is the conditional distribution $p(\mathbf{v}|\mathbf{o}, \mathbf{f})$. Since the true system evolves according to a set of laws that are (presumably) independent of any accessible forecast, $\mathbf{f}$ and $\mathbf{v}$ are conditionally independent, in the sense that

$$p(\mathbf{v}|\mathbf{f}, \mathbf{o}) = p(\mathbf{v}|\mathbf{o}). \tag{4}$$

Hence, if the joint distribution $p(\mathbf{v}, \mathbf{o}, \mathbf{f})$ were really known, then the accessible forecast would be irrelevant for the purposes of measuring predictability. The fact that knowledge of $p(\mathbf{v}, \mathbf{o}, \mathbf{f})$ is tantamount to knowledge of the perfect model distribution $p(\mathbf{v}|\mathbf{o})$ raises the question as to the role of the forecast model. The answer lies in the fact that some distributions are more accessible than others. For instance, the distributions $p(\mathbf{v}, \mathbf{o})$ and $p(\mathbf{f}, \mathbf{o})$ are anticipated to be complicated functions owing to the nonlinear transition probabilities associated with dynamical systems. On the other hand, if the forecast model captures enough detail in the nonlinear processes, then it is hoped that the forecast $\mathbf{f}$ will differ from $\mathbf{v}$ in "simple" ways that are "easily" corrected. For instance, if the forecast merely is biased relative to $\mathbf{v}$, then the best prediction is the forecast distribution shifted by an amount that removes the bias. In this scenario, $p(\mathbf{v}, \mathbf{o})$ and $p(\mathbf{f}, \mathbf{o})$ would be impractical to estimate owing to their nonlinearity, but $p(\mathbf{v}, \mathbf{f})$ would not because $\mathbf{v}$ and $\mathbf{f}$ are related by an additive constant. The approach pursued here assumes that $p(\mathbf{v}, \mathbf{f})$ requires "much less" data for its estimation than $p(\mathbf{v}, \mathbf{o})$, otherwise we would use $p(\mathbf{v}, \mathbf{o})$ and dispense with the forecast altogether. Some insight into these assumptions is provided by the examples presented in section 7.

Since our focus is on predictability, we do not dwell on the question of how to utilize $p(\mathbf{v}, \mathbf{f})$ to characterize forecast errors; see Murphy (1993) and von Storch and Zwiers (1999) for discussion of this issue. Given the joint distribution $p(\mathbf{v}, \mathbf{f})$, the distribution of the verification given the accessible forecast is $p(\mathbf{v}|\mathbf{f})$. We call $p(\mathbf{v}|\mathbf{f})$ the regression forecast distribution for reasons that will become apparent. The regression forecast distribution $p(\mathbf{v}|\mathbf{f})$ has many desirable properties related to accuracy, reliability, and resolution, in the sense of Murphy (1993). Furthermore, if the accessible forecast is independent of the verification, then $p(\mathbf{v}|\mathbf{f}) = p(\mathbf{v})$, the regression forecast distribution, reduces to the climatological distribution and the variable $\mathbf{v}$ is said to be unpredictable. In such cases, the accessible forecast distribution gives no information about the verification that is not already contained in the climatological distribution.

If an ensemble of forecasts are available, say $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M$, then the desired regression forecast distribution is the conditional distribution given the forecast ensemble $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$. Typically, forecast ensembles from the same model are constructed such that each member is equally likely. In such cases, the order of the ensembles is irrelevant and hence the regression forecast distribution $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$ can depend on the sample only through certain sufficient statistics. In section 6 we show that if the distribution is joint normal, then the ensemble mean is a sufficient statistic of the regression forecast distribution; that is, $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M) = p(\mathbf{v}|\langle\mathbf{f}\rangle)$, where $\langle\mathbf{f}\rangle$ is the sample mean of the ensemble forecasts [also defined in (18)].

Note that the forecast distribution $p(\mathbf{f}|\mathbf{o})$ plays a relatively minor role in predictability, as compared to the regression forecast distribution $p(\mathbf{v}|\mathbf{f})$. This point deserves mention since numerous predictability studies focus almost exclusively on the forecast distribution $p(\mathbf{f}|\mathbf{o})$. This emphasis is appropriate if the accessible forecast is perfect. If the accessible forecast is not perfect however, structure in the forecast distribution is relevant only to the extent that it covaries with the event in question. One might suggest that the forecast distribution can be transformed into the relevant distribution for the event that we want to predict. Leaving aside the question of how to construct an appropriate transformation, there can be no more information in the derived forecast distribution than in the individual members of the forecast that were used to construct the distribution. Hence, it is sensible that the regression forecast distribution, which plays a central role in our predictability framework, depends on the actual realizations of the forecast rather than on some distribution derived from the realizations.

4. Predictability of regression forecasts

The previous section introduced the regression forecast distribution $p(\mathbf{v}|\mathbf{f})$, which can be interpreted as the distribution of the future state $\mathbf{v}$ given the accessible forecast $\mathbf{f}$ and the (past) joint behavior between these two variables. This section shows that the predictability of a regression forecast distribution constitutes a rigorous lower bound on the true predictability. Furthermore, the "potential predictability," which is a measure of the difference between the accessible forecast distribution $p(\mathbf{f}|\mathbf{o})$ and its climatology $p(\mathbf{f})$, provides an upper bound on the predictability of the regression forecast distribution. These bounds clarify the role of accessible forecasts in the estimation of predictability.

As discussed in Part I, the most fundamental definition of predictability is based on some measure of the

difference between the perfect model distribution $p(\mathbf{v}|\mathbf{o})$ and climatological distribution $p(\mathbf{v})$. Two measures of this difference are relative entropy $R_{v,o}$ and predictive information $P_{v,o}$, as discussed in Kleeman (2002) and Schneider and Griffies (1999). DelSole (2004b) showed that the average of either of these quantities, over all observations, yields a quantity called the mutual information $I(\mathbf{V}; \mathbf{O})$:

$$I(\mathbf{V}; \mathbf{O}) = \int p(\mathbf{o}) R_{v,o}\, d\mathbf{o} = \int p(\mathbf{o}) P_{v,o}\, d\mathbf{o}. \qquad (5)$$

Mutual information has the attractive property that it is invariant with respect to invertible, nonlinear transformations.

Unfortunately, the perfect model distribution $p(\mathbf{v}|\mathbf{o})$ is not accessible, as discussed in the previous section. Instead, we have access to the forecast $p(\mathbf{f}|\mathbf{o})$ and the regression forecast distribution $p(\mathbf{v}|\mathbf{f})$. Hence, two distinct predictability measures may be conceived. First, the accessible forecast $p(\mathbf{f}|\mathbf{o})$ may be compared to its climatology $p(\mathbf{f})$. By analogy with (5), the associated average predictability is the mutual information between $\mathbf{F}$ and $\mathbf{O}$, denoted $I(\mathbf{F}; \mathbf{O})$. Here $I(\mathbf{F}; \mathbf{O})$ will be called potential predictability since this term is used similarly in the literature to describe the predictability of a forecast system relative to its climatology, without reference to the true system. Second, the regression forecast distribution $p(\mathbf{v}|\mathbf{f})$ may be compared to the climatological distribution $p(\mathbf{v})$. It can be shown that the average predictability of the regression forecast distribution is the mutual information between $\mathbf{F}$ and $\mathbf{V}$, denoted $I(\mathbf{F}; \mathbf{V})$; $I(\mathbf{F}; \mathbf{V})$ will be called the predictability of the regression forecast distribution.

We now show that the metrics $I(\mathbf{V}; \mathbf{O})$, $I(\mathbf{F}; \mathbf{O})$, and $I(\mathbf{F}; \mathbf{V})$ satisfy certain fundamental inequalities. First, Eq. (4) implies that the variables $\mathbf{v}$, $\mathbf{f}$, $\mathbf{o}$ form a Markov chain in the order $\mathbf{f} \Rightarrow \mathbf{o} \Rightarrow \mathbf{v}$. By the fundamental data processing theorem in information theory (Cover and Thomas 1991, chapter 2), the mutual information between the above variables satisfy the inequality

$$I(\mathbf{F}; \mathbf{O}) \geq I(\mathbf{F}; \mathbf{V}). \qquad (6)$$

This inequality states that the potential predictability of an accessible forecast system is greater than or equal to predictability of the regression forecast distribution. The above inequality clarifies that potential predictability does not constitute an upper bound on the true predictability, as is sometimes implied in the literature but, rather, it constitutes an upper bound on the average predictability of the *regression forecast*.

Conditional independence of the verification and accessible forecast also implies the opposite Markov chain $\mathbf{v} \Rightarrow \mathbf{o} \Rightarrow \mathbf{f}$, from which it follows that

$$I(\mathbf{V}; \mathbf{O}) \geq I(\mathbf{F}; \mathbf{V}). \qquad (7)$$

This inequality states that no regression forecast can have greater predictability than that of the true system. Equivalently, the predictability of the regression forecast distribution constitutes a lower bound on the average predictability of the true system.

In contrast to most other proposed measures of predictability, mutual information does not require that the state space of the accessible forecast and the true system be the same. The desirability of this property can be appreciated from the fact that investigators generally are interested in whether some set of forecast variables can provide useful predictors of the verification, regardless of whether the variables are the same. Large mutual information indicates that some forecast variables are statistically dependent with the verification and hence can provide useful predictors of the verification.

A schematic that may facilitate the interpretation of the above quantities is shown in Fig. 1. In this abstraction, a probability distribution is characterized by a "point," and the distance between two points indicates the difference between two distributions. Each line segment in the figure has been labeled to indicate its meaning. Angles have no meaning. The distance between the climatological distribution $p(\mathbf{v})$ and perfect model distribution $p(\mathbf{v}|\mathbf{o})$ defines the predictability of the true system. However, we do not have access to the perfect model distribution $p(\mathbf{v}|\mathbf{o})$, rather, we have access to the accessible forecast distribution $p(\mathbf{f}|\mathbf{o})$. The "distance"

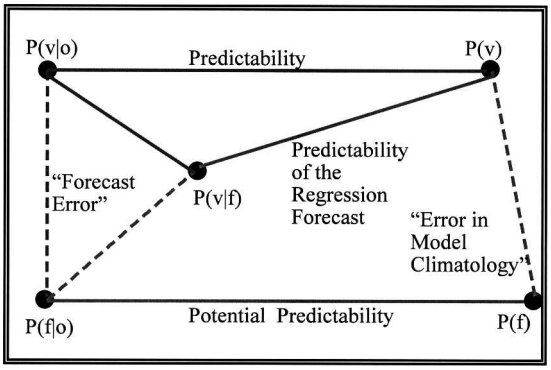

FIG. 1. Schematic illustrating the relation between different distributions and the associated concepts. A distribution is represented by a point and the "distance" between the points indicates the degree of difference between the distributions. Each line segment is labeled to indicate its meaning. The dashed lines join distributions that usually are represented in different state spaces and therefore have undefined distances.

between the perfect model distribution $p(\mathbf{v}|\mathbf{o})$ and accessible forecast $p(\mathbf{f}|\mathbf{o})$ is the most complete description of forecast error (although this distance is undefined if these distributions are represented in different state spaces). The distance between the accessible forecast $p(\mathbf{f}|\mathbf{o})$ and its climatology $p(\mathbf{f})$ represents potential predictability. The regression forecast distribution $p(\mathbf{v}|\mathbf{f})$ provides a link between these two types of predictability measures. The distance between the regression forecast distribution $p(\mathbf{v}|\mathbf{f})$ and the climatological distribution $p(\mathbf{v})$ is the best estimate of predictability based solely on the accessible forecasts. The figure has been constructed such that the predictability of the regression forecast distribution is smaller than either the predictability of the true system or the predictability of the accessible forecast system, as required by inequalities (6) and (7).

A remarkable property of mutual information is that no operation on the forecast can increase mutual information, provided that the operation in question is independent of the verification. To show this, suppose that for a given forecast $\mathbf{f}$ we attempt to construct a new forecast, denoted $L(\mathbf{f})$, with the goal of improving the skill over the original forecast $\mathbf{f}$. If the operation $L(\cdot)$ is a function only of $\mathbf{f}$, then the distribution of $L(\mathbf{f})$, conditional on $\mathbf{f}$ and $\mathbf{v}$, must be independent of $\mathbf{v}$:

$$p(L(\mathbf{f})|\mathbf{f}, \mathbf{v}) = p(L(\mathbf{f})|\mathbf{f}). \tag{8}$$

It follows from this expression that the above variables form a Markov chain in the order

$$\mathbf{v} \Rightarrow \mathbf{f} \Rightarrow L(\mathbf{f}). \tag{9}$$

By the fundamental data processing theorem in information theory (Cover and Thomas 1991, chapter 2), the mutual information between the variables satisfy the inequality

$$I(\mathbf{V}; \mathbf{F}) \geq I(\mathbf{V}; L(\mathbf{F})). \tag{10}$$

Hence, no manipulation of the forecast can enhance $I(\mathbf{V};\mathbf{F})$. This property distinguishes $I(\mathbf{V}; \mathbf{F})$ from other skill metrics, such as mean square error, which often can be improved by biasing the forecast toward climatology. Since mutual information is invariant with respect to invertible, nonlinear transformations, the above proof implies that noninvertible transformations can only reduce mutual information. The above inequality has an intuitive interpretation in communication theory: It states that, if a message is sent through a noisy channel and received at the other end as an output, no manipulation of the output can increase the information about the message contained in the output.

Mutual information between forecast and verification also can be interpreted as a measure of skill, as suggested briefly by Leung and North (1990). The skill of a forecast can be measured in at least two distinct ways: by the "closeness" between forecast and verification, such as as measured by mean square error, or by the "temporal similarity" between forecast and verification, as measured by the correlation coefficient. Mutual information can be interpreted as a generalization of "similarity" measures since it is based on the fundamental probabilistic definition of independence and, hence, does not make implicit assumptions regarding the form of the relation between two variables. Furthermore, mutual information is invariant with respect to nonlinear transformations of the data, is invariant with respect to the role of forecast and verification, cannot be improved by manipulating the forecast (provided the manipulation is independent of verification), and vanishes if and only if the forecast is statistically independent of the verification. Also, owing to (5), forecasts with larger mutual information provide more information about the verification, which is a sensible measure of skill. Finally, in the case of bivariate normal distributions, mutual information is monotonically related to the correlation between forecast and verification. Therefore, mutual information reduces to a common measure of skill in suitable circumstances.

The above discussion implies that a forecast can contain large systematic errors and yet still have significant mutual information. In communication theory, we would say that the forecast is subject to distortion. Such distortion can be corrected if the functional relation between the forecast and verification can be inverted. This correction is implicitly included in the regression forecast distribution $p(\mathbf{v}|\mathbf{f})$. Adopting mutual information as a measure of skill would implicitly include this correction and hence eliminate the temptation to statistically correct forecasts for the purposes of improving skill. These comments should not be construed as suggesting that skill metrics, such as mean square error, are not useful. We are simply clarifying the fact that mutual information measures a different kind of skill than mean square error.

For continuous distributions, mutual information has no maximum value. Joe (1989) showed that the transformation $[1 - \exp(-2I)]^{-1/2}$ produces a value in the interval [0, 1] and recovers the correlation, multiple correlation, and partial correlation in appropriate circumstances when the variables are normally distributed. Schneider and Griffies (1999) propose an analogous transformation for predictive information. For discrete distributions, mutual information is bounded above by the entropy of the verification $H(\mathbf{V})$. Accordingly, in the discrete case, the ratio $I(\mathbf{V}; \mathbf{F})/H(\mathbf{V})$ is bounded between 0 and 1 and, hence, may provide an

attractive skill score for discrete, probabilistic forecast verification. Joe discusses other normalizations of mutual information.

Mutual information can be interpreted not only as a measure of the dependence between variables, but also as a measure of the reduction of uncertainty when one variable becomes known. The latter interpretation follows from the identity

$$I(\mathbf{V}; \mathbf{F}) = H(\mathbf{V}) - H(\mathbf{V}|\mathbf{F}), \tag{11}$$

where $H(\mathbf{V})$ is the entropy of the climatological distribution $p(\mathbf{v})$ and $H(\mathbf{V}|\mathbf{F})$ is the conditional entropy of $\mathbf{V}$ given $\mathbf{F}$ [Cover and Thomas (1991) chapter 2]. According to this identity, positive skill $I(\mathbf{V}; \mathbf{F})$ implies $H(\mathbf{V}) > H(\mathbf{V}|\mathbf{F})$, implying that a skillful forecast reduces the average uncertainty of an event relative to the climatological distribution. This relation links the concepts of degree of dependence (skill) and reduction in uncertainty (predictability). Inequality (7) and identity (11) imply $H(\mathbf{V}|\mathbf{F}) \geq H(\mathbf{V}|\mathbf{O})$, which states that the forecast cannot reduce uncertainty more than observations and a perfect forecast model. The above results can be extended to show that conditioning never increases the average uncertainty. It follows that a forecast based on all available knowledge should have less uncertainty than a forecast based on partial knowledge.

5. Predictable components of an accessible forecast

An unpredictable component of a forecast is a random variable $\mathbf{F}_u$ that satisfies

$$p(\mathbf{f}_u|\mathbf{o}) = p(\mathbf{f}_u). \tag{12}$$

If a forecast variable does not satisfy (12), then it is called a potential predictable component, denoted $\mathbf{F}_p$. The word potential is used to indicate that these components are predictable in the accessible forecast but not necessarily in the true system, though this term often will be dropped in sequel because predictable components of other forecasts will not be considered. In this section, we show that, under certain plausible assumptions, potential predictable components, and these components alone, can be used as predictors of a regression forecast. This result has important implications if the potential predictable components span a dimension smaller than that of the full system.

The proof given below holds even if the variables are not joint normally distributed. In practice, however, the normal assumption is needed to identify predictable components. For instance, if the variables are joint normally distributed, then canonical correlation analysis (CCA) can identify the unpredictable components (DelSole 2004b). This procedure is equivalent to predictable component analysis proposed by Schneider and Griffies (1999), provided the distributions are joint normal (DelSole and Chang 2003). Even if variables are not normally distributed, CCA still might be a useful method of finding potential predictable components because it identifies components with large correlation. Whether more general methods of finding predictable components are needed for realistic systems is a question that only experiment can settle.

Suppose the forecast variables can be split into two groups: those with vanishing mutual information, $\mathbf{Z}_u = \{\mathbf{f}_u^{(1)}, \mathbf{f}_u^{(2)}, \ldots\}$, called potential unpredictable components, and everything else, $\mathbf{Z}_p = \{\mathbf{f}_p^{(1)}, \mathbf{f}_p^{(2)}, \ldots\}$, called potential predictable components. The unpredictable components are identified with weather noise. As such, it is plausible to assume that the unpredictable components are independent of observations jointly:

$$p(\mathbf{Z}_u|\mathbf{o}) = p(\mathbf{Z}_u). \tag{13}$$

We make the stronger, yet still plausible, assumption that weather noise in the accessible forecast is independent of observations, verification, and predictable components:

$$p(\mathbf{Z}_u|\mathbf{o}, \mathbf{v}, \mathbf{Z}_p) = (\mathbf{Z}_u). \tag{14}$$

This assumption holds automatically if weather noise is parameterized as independent, additive noise, as is usually the case in predictability studies. Note that the above assumption implies that $\mathbf{Z}_u$ is independent of any combination of $\mathbf{o}$, $\mathbf{v}$, $\mathbf{Z}_p$.

Assumption (14) implies that the regression forecast distribution can be expressed as

$$\begin{aligned} p(\mathbf{v}|\mathbf{Z}_p, \mathbf{Z}_u) &= \frac{p(\mathbf{v}, \mathbf{Z}_p, \mathbf{Z}_u)}{p(\mathbf{Z}_p, \mathbf{Z}_u)} \\ &= \frac{p(\mathbf{Z}_u|\mathbf{Z}_p, \mathbf{v})p(\mathbf{Z}_p, \mathbf{v})}{p(\mathbf{Z}_u|\mathbf{Z}_p)p(\mathbf{Z}_p)} \\ &= \frac{p(\mathbf{Z}_u)p(\mathbf{Z}_p, \mathbf{v})}{p(\mathbf{Z}_u)p(\mathbf{Z}_p)} \\ &= p(\mathbf{v}|\mathbf{Z}_p). \end{aligned} \tag{15}$$

It follows immediately from this identity that

$$I(\mathbf{V}; \mathbf{Z}) = I(\mathbf{V}; \mathbf{Z}_p). \tag{16}$$

The importance of the above identities, (15) and (16), is that the dimension of $\mathbf{Z}_p$ may be much smaller than the dimension of full system, especially in the context of monthly or seasonal predictability. Furthermore, the predictable components of an accessible forecast model can be determined with more accuracy than the predictable components of the observed system, owing to

the availability of multiple realizations of the accessible forecast. Finally, inequality (6) implies $I(\mathbf{Z}_p; \mathbf{O}) \geq I(\mathbf{V}; \mathbf{Z}_p)$: The predictability of the potential predictable components is never less than the predictability of the regression forecast distribution. For these reasons, predictable components may provide an attractive basis set for reducing the dimension of the predictability analysis.

6. Regression forecasts for Gaussian, Markov systems

In this section, we illustrate the above concepts in the context of stationary, Gaussian, Markov systems. Before doing this, it is instructive to write expressions for the above quantities for joint normally distributed variables. If **v** and **f** are joint normally distributed, then it is well known that the conditional distribution $p(\mathbf{v}|\mathbf{f})$ is

$$p(\mathbf{v}|\mathbf{f}) \sim N_p(\boldsymbol{\mu}_v + \boldsymbol{\Sigma}_{vf}(\boldsymbol{\Sigma}_f^{\tau})^{-1}(\mathbf{f} - \boldsymbol{\mu}_f), \boldsymbol{\Sigma}_v - \boldsymbol{\Sigma}_{vf}(\boldsymbol{\Sigma}_f^{\tau})^{-1}\boldsymbol{\Sigma}_{fv}), \quad (17)$$

where $\boldsymbol{\mu}_v$ and $\boldsymbol{\Sigma}_v$ are the mean and covariance matrix of the marginal distribution p(**v**), $\boldsymbol{\mu}_f$ and $\boldsymbol{\Sigma}_f^{\tau}$ are the analogous quantities for p(**f**), and $\boldsymbol{\Sigma}_{vf} = \boldsymbol{\Sigma}_{fv}^T$ is the cross-covariance matrix between **f** and **v** [see Johnson and Wichern (1982), p. 170 for a derivation]. Readers familiar with statistical methods will recognize this result as equivalent to a least squares linear prediction of **v,** given **f**, for asymptotically large sample size. Although least squares estimation would be an obvious way of correcting a forecast, it arises here not because we are trying to minimize forecast error variance directly, but because, as is well known, it is equivalent to the conditional distribution $p(\mathbf{v}|\mathbf{f})$ for Gaussian variables.

The above considerations assume the availability of a single forecast. Now consider an ensemble of forecasts drawn randomly from $p(\mathbf{f}|\mathbf{o})$. Let the forecast ensemble be $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M$. The conditional distribution of the verification given the forecast ensemble is $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$. For Gaussian variables, the distribution $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$ is identical to $p(\mathbf{v}|\langle \mathbf{f} \rangle)$, where $\langle \mathbf{f} \rangle$ is the sample ensemble mean forecast

$$\langle \mathbf{f} \rangle = \frac{1}{M} \sum_{k=1}^{M} \mathbf{f}_k. \quad (18)$$

The equivalence between $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$ and $p(\mathbf{v}|\langle \mathbf{f} \rangle)$ can be seen in several ways. Perhaps the simplest is to note that, owing to the Gaussian form, the distribution $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$ can depend only linearly with respect to the forecasts $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M$. Furthermore, since there is no basis for treating any one forecast from the same model differently from the others, the distribution $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$ must be invariant with respect to an interchange of any two forecasts. The properties of invariance and linearity imply that the distribution $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$ can depend only on the sum over M vectors $\mathbf{f}_1 + \mathbf{f}_2 + \ldots + \mathbf{f}_M$, which is proportional to the sample ensemble mean $\langle \mathbf{f} \rangle$. From the joint normal distribution assumption, it follows that the conditional distribution $p(\mathbf{v}|\langle \mathbf{f} \rangle)$ is

$$p(\mathbf{v}|\langle \mathbf{f} \rangle) \sim N_p(\boldsymbol{\mu}_v + \boldsymbol{\Sigma}_{v\langle f \rangle}\boldsymbol{\Sigma}_{\langle f \rangle}^{-1}(\langle \mathbf{f} \rangle - \boldsymbol{\mu}_{\langle f \rangle}), \boldsymbol{\Sigma}_v - \boldsymbol{\Sigma}_{v\langle f \rangle}\boldsymbol{\Sigma}_{\langle f \rangle}^{-1}\boldsymbol{\Sigma}_{\langle f \rangle v}), \quad (19)$$

which has the same form as (17), but with mean $\boldsymbol{\mu}_{\langle f \rangle}$ and covariance matrices $\boldsymbol{\Sigma}_{v\langle f \rangle}$ and $\boldsymbol{\Sigma}_{\langle f \rangle}$ pertaining to the sample ensemble mean forecast $\langle \mathbf{f} \rangle$.

Since the regression forecast distribution $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$ depends only on the sample mean forecast $\langle \mathbf{f} \rangle$, it might appear that the regression forecast distribution is independent of the forecast spread. This is not the case, as we will now demonstrate. The population mean forecast, often called the signal, is a random variable given by

$$E[\mathbf{f}|\mathbf{o}] = \int \mathbf{f} p(\mathbf{f}|\mathbf{o})\, d\mathbf{f}. \quad (20)$$

Note that E[**f**|**o**] is a random function since it depends on **o**. It is routine to show that

$$\boldsymbol{\Sigma}_f = \boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_n, \quad (21)$$

where

$$\boldsymbol{\Sigma}_s = E[(E[\mathbf{f}|\mathbf{o}] - E[\mathbf{f}])(E[\mathbf{f}|\mathbf{o}] - E[\mathbf{f}])^T]$$
$$\boldsymbol{\Sigma}_n = E[(\mathbf{f} - E[\mathbf{f}|\mathbf{o}])(\mathbf{f} - E[\mathbf{f}|\mathbf{o}])^T], \quad (22)$$

in which $E[]$ with no conditioning represents the expectation over $p(\mathbf{v}, \mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M, \mathbf{o})$. The term $\boldsymbol{\Sigma}_s$ measures the variance of the "signal," while the term $\boldsymbol{\Sigma}_n$ measures the variance of "forecast spread" or "noise." Elementary sampling theory shows that

$$\boldsymbol{\Sigma}_{v\langle f \rangle} = \boldsymbol{\Sigma}_{vf} = \boldsymbol{\Sigma}_{v,E[f|o]}$$
$$\boldsymbol{\Sigma}_{\langle f \rangle} = \boldsymbol{\Sigma}_s + \frac{1}{M}\boldsymbol{\Sigma}_n, \quad (23)$$

where $\boldsymbol{\Sigma}_{\langle f \rangle}$ is the covariance matrix of $\langle \mathbf{f} \rangle$, and $\boldsymbol{\Sigma}_{v\langle f \rangle}$ is the covariance matrix between **v** and $\langle \mathbf{f} \rangle$. These expressions show that $\Sigma_{\langle f \rangle}$ depends on the spread of the forecast $\boldsymbol{\Sigma}_n$. Since $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Sigma}_n$ are positive definite, increasing $\boldsymbol{\Sigma}_n$ increases the variance of $\langle \mathbf{f} \rangle$ but does not alter the covariance $\boldsymbol{\Sigma}_{v\langle f \rangle}$. Thus, appearances to the contrary, the distribution $p(\mathbf{v}|\langle \mathbf{f} \rangle)$ depends on forecast spread in the following sense: given two forecasts with the same sig-

nal but different ensemble spreads, the forecast with larger spread gives rise to a regression forecast distribution with larger uncertainty. The variation of the regression forecast distribution depends on the sample only through the sample ensemble mean.

It is instructive to consider the case of a perfect accessible forecast. In a perfect model scenario, $p(\mathbf{v}|\mathbf{o}) = p(\mathbf{f}|\mathbf{o})$, which implies that

$$\begin{aligned}\boldsymbol{\mu}_v &= \boldsymbol{\mu}_f = \boldsymbol{\mu}_{\langle f\rangle}\\ \boldsymbol{\Sigma}_v &= \boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_n\\ \boldsymbol{\Sigma}_{vf} &= \boldsymbol{\Sigma}_{v\langle f\rangle} = \boldsymbol{\Sigma}_s.\end{aligned} \tag{24}$$

The last relation arises because the forecast and verification can be represented each as a sum of a common signal plus independent noise, and all cross-covariances involving the noise terms vanish. Substituting these expressions into (19) gives a regression forecast distribution $p(\mathbf{v}|\langle\mathbf{f}\rangle)$ that is multivariate Gaussian with mean and covariance matrix

$$\begin{aligned}\boldsymbol{\mu}_{v|\langle f\rangle} &\overset{\text{PM}}{=} \boldsymbol{\mu}_v + \boldsymbol{\Sigma}_s\left(\boldsymbol{\Sigma}_s + \frac{1}{M}\boldsymbol{\Sigma}_n\right)^{-1}(\mathbf{s} - \boldsymbol{\mu}_v)\\ \boldsymbol{\Sigma}_{v|\langle f\rangle} &\overset{\text{PM}}{=} \boldsymbol{\Sigma}_n + \boldsymbol{\Sigma}_s - \boldsymbol{\Sigma}_s\left(\boldsymbol{\Sigma}_s + \frac{1}{M}\boldsymbol{\Sigma}_n\right)^{-1}\boldsymbol{\Sigma}_s,\end{aligned} \tag{25}$$

where "PM" indicates perfect model. In the limit $M \to \infty$, the regression forecast distribution approaches $N(E[\mathbf{f}|\mathbf{o}], \boldsymbol{\Sigma}_n)$, which is the (correct) perfect model distribution $p(\mathbf{v}|\mathbf{o})$. For finite ensemble size M, however, the conditional distribution $p(\mathbf{v}|\langle\mathbf{f}\rangle)$ differs from the perfect model forecast distribution $p(\mathbf{v}|\mathbf{o})$, even for a perfect model scenario, reflecting the fact that the forecast distribution has not been adequately sampled.

Now we consider the evolution of predictability in stationary, Gaussian, Markov systems. As discussed in Part I, the state $\mathbf{x}_t$ of such a system can be interpreted as a solution of a linear stochastic model of the form

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x} + \mathbf{w}, \tag{26}$$

where $\mathbf{A}$ is a stable dynamical operator and $\mathbf{w}$ is a Gaussian white noise process with zero mean and covariance matrix $\mathbf{Q}$. The properties of this system have been discussed extensively in the literature (Gardiner 1990; DelSole 2004b, and references therein). The main facts of relevance in this paper are the following. Assuming the solution to (26) was begun in the infinite past, then $\mathbf{x}_t$ is stationary. If the initial condition is drawn randomly from the stationary distribution $p(\mathbf{x}_t)$, then the marginal distribution for the initial condition $\mathbf{i} = \mathbf{x}_t$ and verification $\mathbf{v} = \mathbf{x}_{t+\tau}$ are equal, independent of time, and normally distributed with zero mean and constant covariance matrix $\boldsymbol{\Sigma}_v$. Thus

$$p(\mathbf{i}) = p(\mathbf{v}) = N(\mathbf{0}, \boldsymbol{\Sigma}_v). \tag{27}$$

The solution to (26) with initial condition $\mathbf{i}$ can be written as the sum of two terms,

$$\mathbf{v} = \mathbf{P}\mathbf{i} + \mathbf{e}_v, \tag{28}$$

where $\mathbf{P} = \exp(\mathbf{A}\tau)$ is the propagator of the system and $\mathbf{e}_v$ is Gaussian white noise with distribution

$$p(\mathbf{e}_v) = N(\mathbf{0}, \boldsymbol{\Sigma}_v - \mathbf{P}\boldsymbol{\Sigma}_v\mathbf{P}^{\mathrm{T}}). \tag{29}$$

The random variables $\mathbf{i}$ and $\mathbf{e}_v$ are independent. It follows from the above two equations that the conditional distribution of **v,** given **i,** is

$$p(\mathbf{v}|\mathbf{i}) = N(\mathbf{P}\mathbf{i}, \boldsymbol{\Sigma}_v - \mathbf{P}\boldsymbol{\Sigma}_v\mathbf{P}^{\mathrm{T}}). \tag{30}$$

The predictability of this system, as measured by relative entropy, predictive information, and mutual information, has been discussed in Part I and need not be reproduced here.

We attempt to forecast $\mathbf{v}$ given $\mathbf{o}$. In this attempt, it would be unrealistic to assume that (26) is perfectly known. Thus, a forecast based on a stochastic model,

$$\frac{d\mathbf{y}}{dt} = \mathbf{A}_f\mathbf{y} + \mathbf{w}_f, \tag{31}$$

is attempted, where $\mathbf{A}_f$ differs from $\mathbf{A}$ in (26), and $\mathbf{w}_f$ is Gaussian white noise with statistics possibly different from those of $\mathbf{w}$ in (26). The propagator of the forecast system is $\mathbf{P}_f = \exp(\mathbf{A}_f\tau)$, and the covariance matrix of the asymptotic forecast is $\boldsymbol{\Sigma}_f^{\infty}$. To this model corresponds an analysis $p(\mathbf{i}_f|\mathbf{o})$, which represents the distribution of the initial condition appropriate for the forecast model. Note that $p(\mathbf{i}_f|\mathbf{o})$ is not equal to $p(\mathbf{i}|\mathbf{o})$, the analysis for the perfect model does not equal the analysis for the accessible forecast model. A forecast by the accessible forecast model starting from $\mathbf{i}_f$ satisfies the equation

$$\mathbf{f} = \mathbf{P}_f\mathbf{i}_f + \mathbf{e}_f, \tag{32}$$

where $\mathbf{e}_f$ is a Gaussian random process with distribution

$$p(\mathbf{e}_f) = N(\mathbf{0}, \boldsymbol{\Sigma}_f^{\infty} - \mathbf{P}_f\boldsymbol{\Sigma}_f^{\infty}\mathbf{P}_f^{\mathrm{T}}). \tag{33}$$

The variables $\mathbf{i}$ and $\mathbf{e}_f$ are independent. Physically, this independence follows from the fact that $\mathbf{i}$ is a realization from the true system (26) whereas $\mathbf{e}_f$ represents the internal noise of the forecast. We could allow $\mathbf{e}_f$ to have nonzero mean, in which case the forecast $\mathbf{f}$ would be biased, but this situation represents only a trivial exten-

sion of the unbiased case. It thus follows from (32) and (33) that

$$p(\mathbf{f}|\mathbf{i}_f) = N(\mathbf{P}_f\mathbf{i}_f, \boldsymbol{\Sigma}_f^{\infty} - \mathbf{P}_f\boldsymbol{\Sigma}_f^{\infty}\mathbf{P}_f^{\mathrm{T}}). \tag{34}$$

Further progress requires clarifying the relation between $\mathbf{i}$, $\mathbf{i}_f$, and $\mathbf{o}$. Since $\mathbf{i}$ and $\mathbf{i}_f$ arise from an analysis procedure, they depend only on the antecedent observations and forecast models. In particular, $\mathbf{i}$ and $\mathbf{i}_f$ are conditionally independent given the observations:

$$p(\mathbf{i}, \mathbf{i}_f|\mathbf{o}) = p(\mathbf{i}|\mathbf{o})p(\mathbf{i}_f|\mathbf{o}). \tag{35}$$

We consider the case in which the initial condition errors are small. Thus, for simplicity, we assume $\mathbf{i} = \mathbf{i}_f$, in which case the triplet ($\mathbf{e}_v$, $\mathbf{e}_f$, $\mathbf{i}$) forms a mutually independent set. Since all three variables are independent and normally distributed, any linear combination of the triple is joint normally distributed. In particular, the pair ($\mathbf{e}_v - \mathbf{P}$ i) and ($\mathbf{e}_f - \mathbf{P}_f\,\mathbf{i}$) are joint normally distributed, from which it follows that $\mathbf{v}$ and $\mathbf{f}$ are joint normally distributed. Thus, the regression forecast distribution can be written immediately as (17), where it remains to determine the covariance matrices (the means are assumed to vanish).

From (34), the marginal distribution $p(\mathbf{f})$ is Gaussian with zero mean and covariance

$$\begin{aligned}\boldsymbol{\Sigma}_f^{\tau} &= E[(\mathbf{P}_f\mathbf{i} + \mathbf{e}_f)(\mathbf{P}_f\mathbf{i} + \mathbf{e}_f)^{\mathrm{T}}] \\ &= \mathbf{P}_f\boldsymbol{\Sigma}_v\mathbf{P}_f^{\mathrm{T}} + \{\boldsymbol{\Sigma}_f^{\infty} - \mathbf{P}_f\boldsymbol{\Sigma}_f^{\infty}\mathbf{P}_f^{\mathrm{T}}\},\end{aligned} \tag{36}$$

where we have used the fact that $\mathbf{i}$ and $\mathbf{e}_f$ are independent. In general, $\boldsymbol{\Sigma}_f^{\tau}$ can depend on lead time. This dependence arises whenever the initial condition for the forecast model is drawn from a distribution that differs from the marginal distribution of the forecast. If $\boldsymbol{\Sigma}_f^{\infty}$ and $\boldsymbol{\Sigma}_v$ are equal, implying that the climatology of the forecast and true system coincide, then the covariance matrix $\boldsymbol{\Sigma}_f^{\tau}$ is independent of τ. The cross-covariance $\boldsymbol{\Sigma}_{vf}$ is

$$\begin{aligned}\boldsymbol{\Sigma}_{vf} &= E[\mathbf{v}\mathbf{f}^{\mathrm{T}}] \\ &= E[(\mathbf{P}\mathbf{i} + \mathbf{e}_v)(\mathbf{P}_f\mathbf{i} + \mathbf{e}_f)^{\mathrm{T}}] \\ &= \mathbf{P}E[\mathbf{i}\mathbf{i}^{\mathrm{T}}]\mathbf{P}_f^{\mathrm{T}} \\ &= \mathbf{P}\boldsymbol{\Sigma}_v\mathbf{P}_f^{\mathrm{T}}.\end{aligned} \tag{37}$$

All quantities in (17) have now been specified. This completes the task of finding the regression forecast distribution $p(\mathbf{v}|\mathbf{f})$ for a stationary, Gaussian, Markov process.

Now consider the more general case of an ensemble of forecasts. An ensemble of forecasts can be interpreted as a set of independent vectors $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M$ drawn randomly from the accessible forecast distribution $p(\mathbf{f}|\mathbf{i})$. As discussed in section 3, the appropriate regression forecast distribution is $p(\mathbf{v}|\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M)$, which in the case of Gaussian distributions is identical to $p(\mathbf{v}|\langle\mathbf{f}\rangle)$, where $\langle\mathbf{f}\rangle$ is the ensemble mean forecast

$$\langle\mathbf{f}\rangle = \frac{1}{M}\sum_{k=1}^{M}\mathbf{f}_k. \tag{38}$$

To derive an expression for $p(\mathbf{v}|\langle\mathbf{f}\rangle)$, we may follow precisely the same procedure used to derive $p(\mathbf{v}|\mathbf{f})$, but with the new variable

$$\mathbf{e}_{\langle f\rangle} = \langle\mathbf{f}\rangle - \mathbf{P}_f\mathbf{i} \tag{39}$$

with distribution

$$p(\mathbf{e}_{\langle f\rangle}|\mathbf{i}) = N\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_f^{\infty} - \mathbf{P}_f\boldsymbol{\Sigma}_f^{\infty}\mathbf{P}_f^{\mathrm{T}}}{M}\right) = p(\mathbf{e}_{\langle f\rangle}). \tag{40}$$

The resulting regression forecast distribution for the ensemble is

$$p(\mathbf{v}|\langle\mathbf{f}\rangle) \sim N(\boldsymbol{\Sigma}_{v\langle f\rangle}\boldsymbol{\Sigma}_{\langle f\rangle}^{-1}\langle\mathbf{f}\rangle, \boldsymbol{\Sigma}_v - \boldsymbol{\Sigma}_{v\langle f\rangle}\boldsymbol{\Sigma}_{\langle f\rangle}^{-1}\boldsymbol{\Sigma}_{\langle f\rangle v}). \tag{41}$$

Standard sampling theory gives

$$\begin{aligned}\boldsymbol{\Sigma}_{\langle f\rangle}^{\tau} &= \mathbf{P}_f\boldsymbol{\Sigma}_v\mathbf{P}_f^{\mathrm{T}} + \frac{1}{M}(\boldsymbol{\Sigma}_f^{\infty} - \mathbf{P}_f\boldsymbol{\Sigma}_f^{\infty}\mathbf{P}_f^{\mathrm{T}}) \\ \boldsymbol{\Sigma}_{v\langle f\rangle} &= \mathbf{P}\boldsymbol{\Sigma}_v\mathbf{P}_f^{\mathrm{T}}.\end{aligned} \tag{42}$$

We recover the covariances for a single realization $\mathbf{f}$, (36) and (37), by substituting $M = 1$ into the above expression. It can be verified that, in the limit of infinite ensemble size $M \to \infty$, the regression forecast distribution (41) asymptotically approaches the perfect model distribution (30), provided $\boldsymbol{\Sigma}_{\langle f\rangle}$ and $\boldsymbol{\Sigma}_{v\langle f\rangle}$ remain nonsingular.

The predictive information, mutual information, and relative entropy for the regression forecast distribution p($\mathbf{v}|\langle\mathbf{f}\rangle$) are obtained by substituting (41) into the appropriate expressions in Part I. The results are

$$\begin{aligned}P_{\langle f\rangle} &= -\frac{1}{2}\log|\mathbf{I} - \mathbf{Z}\mathbf{Z}^{\mathrm{T}}| \\ I_{\langle f\rangle} &= -\frac{1}{2}\log|\mathbf{I} - \mathbf{Z}\mathbf{Z}^{\mathrm{T}}| \\ R_{\langle f\rangle} &= -\frac{1}{2}\log|\mathbf{I} - \mathbf{Z}\mathbf{Z}^{\mathrm{T}}| + \frac{1}{2}Tr\{\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\} \\ &\quad + \frac{1}{2}\langle\mathbf{f}^{\mathrm{T}}\rangle\boldsymbol{\Sigma}_{\langle f\rangle}^{-1/2}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})\boldsymbol{\Sigma}_{\langle f\rangle}^{-1/2}\langle\mathbf{f}\rangle,\end{aligned} \tag{43}$$

where

$$\mathbf{Z} = \boldsymbol{\Sigma}_v^{-1/2}\boldsymbol{\Sigma}_{v\langle f\rangle}\boldsymbol{\Sigma}_{\langle f\rangle}^{-1/2}. \tag{44}$$

These expressions are isomorphic to those obtained for the perfect model distributions and, hence, have prop-

erties similar to those associated with the true system (i.e., relative entropy depends on initial condition, all three quantities decay monotonically with lead time, etc.).

7. Numerical examples

We now give numerical examples to illustrate the above concepts. Consider first a two dimensional system with noise covariance matrix $\mathbf{Q} = \mathbf{I}$, and dynamical operator

$$\mathbf{A} = \begin{pmatrix} -1/5 & \beta \\ 0 & -1 \end{pmatrix}, \tag{45}$$

where β is a tunable parameter. This model can be solved analytically by methods described in Gardiner (1990) and DelSole (2004b). Since $\mathbf{A}$ is upper triangular, its eigenvalues are $-1/5$ and -1, regardless of β. The case $\beta = 0$ corresponds to a normal dynamical operator in prewhitened coordinates, which constitutes a lower bound on the predictability of all stochastic systems with the same eigenvalues (Tippett and Chang 2003). Suppose that the "truth" is identified with the stochastic model with $\beta = 0$, while the accessible forecast is identified with $\beta = 5$; in both cases the covariance matrix is assumed to be $\mathbf{Q} = \mathbf{I}$. This experiment may be termed a perfect initial condition scenario since uncertainty arises from stochastic forcing within the model and not from the initial condition. The mutual information between verification and initial condition in the true system is given by

$$I(\mathbf{v}, \mathbf{i}) = -\frac{1}{2}\log|\mathbf{I} - \mathbf{\Sigma}_{vi}\mathbf{\Sigma}_o^{-1}\mathbf{\Sigma}_{iv}\mathbf{\Sigma}_v^{-1}|, \tag{46}$$

and is shown as the dashed line in Fig. 2. The solid lines show, for different ensemble sizes, the mutual information between verification and accessible forecast, as evaluated from $\mathbf{I}_{\langle \mathbf{f} \rangle}$ in (43). First, note that the predictability of the regression forecast distribution is always less than or equal to the predictability of the system. This reflects the inequality (7), which states that the predictability of a regression forecast distribution is a lower bound on the predictability of the true system. Second, note that the gain in predictability due to doubling the ensemble size is modest after one time unit. This is not surprising given that the regression forecast distribution of joint normal distributions depends on the sample only through the ensemble mean, so extra ensembles merely refine the sample mean. The predictability of the regression forecast distribution converges to the true predictability more rapidly as the β in the forecast model approaches the true value of β.

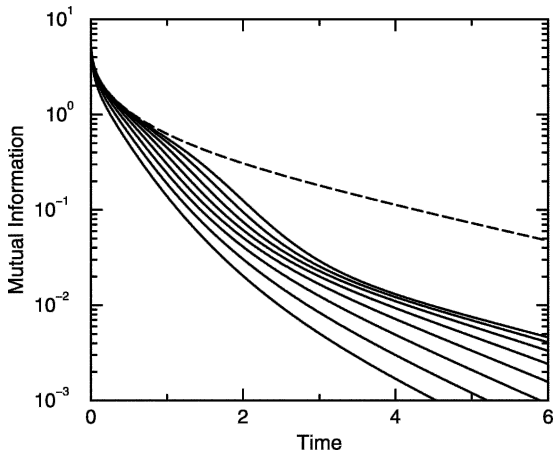

FIG. 2. Predictability of the two-variable stochastic model (26) with dynamical operator (45) and $\beta = 0$ (dashed), and of the regression forecast distribution with $\beta = 5$ for ensemble sizes 1, 2, 4, 8, 16, 32, 64, and 128 (solid lines, from bottom up).

How well can mutual information be estimated from finite samples? To gain insight into this question, we numerically generated time series from the above stochastic models using a forward Euler stochastic scheme (Kloeden and Platen 1999, p. 305) with a time step of 0.01 time units. The verification and observation time series were constructed by first integrating the true stochastic model (i.e., $\beta = 0$) and then sampling this single realization 10 times every 16 time units. Since the slowest decaying eigenmode has an e-folding time of 5 time units, sampling every 16 time units ensures that each verification–initial condition pair is effectively independent of all other such pairs. Within each 16 time unit interval, the initial condition is identified with the starting point and the verification is the value of the time series τ time units later. Accessible forecasts were constructed by starting at each initial condition and integrating the stochastic model using $\beta = 5$, using random forcing that was independent of that used to generate the truth. Multiple ensemble members were generated by integrating from the same initial condition but with independent realizations of the random forcing. The result of this procedure is to produce 10 **v**–**i** pairs and 10 **v**–$\langle$**f**$\rangle$ pairs. Estimates of $I(\mathbf{v}, \mathbf{i})$ and $I(\mathbf{v}, \mathbf{f})$, denoted $I_e(\mathbf{v}, \mathbf{i})$ and $I_e(\mathbf{v}, \mathbf{f})$, were obtained by replacing population covariance matrices with sample covariance matrices in (43) and (46), respectively. This procedure was repeated 500 times to estimate the distribution of $I_e(\mathbf{v}, \mathbf{i})$ and $I_e(\mathbf{v}, \mathbf{f})$.

Figure 3 reproduces the exact values of $I(\mathbf{v}, \mathbf{i})$ and $I(\mathbf{v}, \langle \mathbf{f} \rangle)$ for this model for 16 ensemble members (dashed and solid lines, with no filled circles, respectively). The figure also shows the mean and the 10th and 90th per-

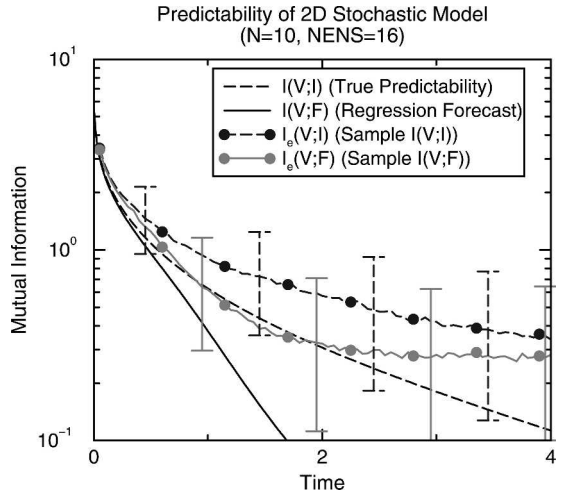

FIG. 3. Predictability of the two-variable stochastic model (26) with $\beta = 0$ (dash with no error bar or circle) and associated regression forecast distribution with $\beta = 5$ and with 16 ensemble members (solid with no error bar or circle), as in Fig. 2. Also shown is the predictability estimated from 10 independent samples of the verification and initial condition (dash with error bars and circles) and associated regression forecast distribution (solid with error bars and circles), based on a Gaussian assumption for the distributions. The error bars show the 10th and 90th percentiles.

centiles, as error bars, of the sample estimates $I_e(\mathbf{v}, \mathbf{i})$ and $I_e(\mathbf{v}, \langle\mathbf{f}\rangle)$ (dashed and solid lines, with error bars). First, we see that the sample estimates $I_e(\mathbf{v}, \mathbf{i})$ and $I_e(\mathbf{v}, \langle\mathbf{f}\rangle)$ tend to be biased upward relative to their respective exact values $I(\mathbf{v}, \mathbf{i})$ and $I(\mathbf{v}, \langle\mathbf{f}\rangle)$. The magnitude of this bias decreases as the sample size increases. Second, we see that $I_e(\mathbf{v}, \mathbf{i})$ tends to be larger than $I_e(\mathbf{v}, \langle\mathbf{f}\rangle)$, even thought the latter is computed from 16 ensemble members. Quantitatively, $I_e(\mathbf{v}, \mathbf{i}) > I_e(\mathbf{v}, \langle\mathbf{f}\rangle)$ in over 75% of the results for time lags less than 3 time units. We have verified that this result holds even if the accessible forecast model is perfect (i.e., if the accessible forecast model has $\beta = 0$, but with random forcing that is independent of the truth), provided the number of forecast ensemble members is less than 10. These limited results suggest that, if the system is joint normally distributed, there is no compelling reason to utilize accessible forecasts, since higher (but equally biased) estimates of predictability can be obtained by estimating $I(\mathbf{v}, \mathbf{i})$ directly from the observed time series. Conceivably, prior knowledge of the system could be incorporated into the estimation procedure to improve the predictability estimates, but this was not explored.

Now consider the nonlinear dynamical model of Lorenz (1963) with parameter values $\sigma = 10$, $\rho = 8/3$, and $\beta = 28$, for which the model is chaotic. This model is distinguished from the previous model in that it is nonlinear and non-Gaussian. This model was integrated with a fourth-order Runge–Kutta scheme starting from a random point near the origin to construct a single time series of length 2600 time units. After computing this time series, the initial 1000 time units were discarded to avoid spinup effects, and independent random numbers from a Gaussian distribution with zero mean and unit variance were added to the time series. The resulting time series then was sampled every 16 time units to construct 100 initial conditions. The accessible forecasts were constructed by integrating the Lorenz model at each of the 100 initial conditions, but with $\beta = 20$, all other parameters held the same (our major conclusions below do not appear to depend on the parameter being perturbed). Additional initial conditions for generating ensemble members were constructed by adding new, independent random numbers to the original solution to the Lorenz model. The resulting 100 **v**–**i** pairs and 100 **v**–$\langle\mathbf{f}\rangle$ pairs are insufficient to estimate distributions along the lines of Kleeman (2002). Nevertheless, the sample size is typical in climate research.

Given the small sample size, we evaluated mutual information by (incorrectly) assuming a Gaussian form for the distributions so that the Eqs. (43) and (46) can be used. Figure 4 shows $I_e(\mathbf{v}, \mathbf{i})$ (dashed) and $I_e(\mathbf{v}, \mathbf{f})$ (solid) estimated from the time series for one ensemble member. We see that $I_e(\mathbf{v}, \mathbf{f})$ exceeds $I_e(\mathbf{v}, \mathbf{i})$, in contrast

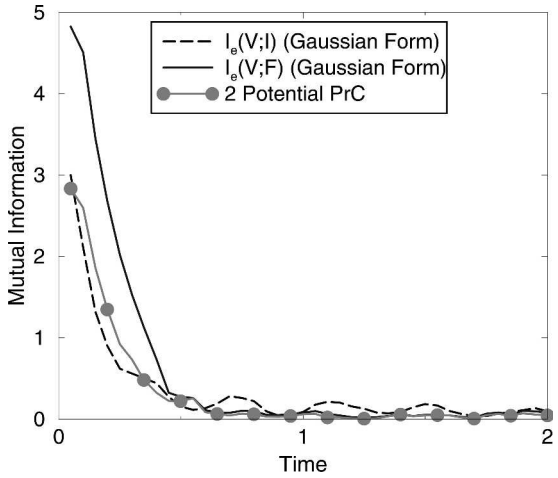

FIG. 4. Predictability of the Lorenz (1963) model estimated from 100 independent initial condition and verification pairs, assuming a Gaussian form for the distribution (dashed). The accessible forecast model is drawn from the same model class but with the parameter β adjusted from 28 to 20. The predictability of the regression forecast distribution based on one ensemble member and a Gaussian form for the distribution is shown as the solid line. The predictability based only on two leading potential predictable components of the accessible forecast model is shown as the solid line with circles.

to Fig. 3. This result does not contradict inequality (7), which pertains to the exact probability distributions, because here a Gaussian form is imposed for the distributions. The essential reason for this result is that the accessible forecasts track the verification better than a linear prediction based on the initial condition, presumably because the accessible forecast to some extent captures important nonlinearity. Consequently, the covariance between $\mathbf{v}$ and $\mathbf{f}$ remains much higher as τ increases than the covariance between $\mathbf{v}$ and $\mathbf{i}$. Interestingly, adding additional ensemble members does not substantially improve estimates of mutual information. The reason for this is that each ensemble member remains relatively close to the other members over the time scales considered; that is, each $\mathbf{f}$ is close to $\langle \mathbf{f} \rangle$. Thus, adding new ensemble members does not add much information about the prediction.

The potential predictable components were obtained by performing CCA between 100 **f–i** pairs. The mutual information between $\mathbf{v}$ and the first two predictable components $\mathbf{f}_p$, as computed from $I_{\langle f \rangle}$ in (43), is shown in Fig. 4 as the line with circles. Although the curve shows that the predictability based on two predictable components is comparable to the (Gaussian) mutual information between $\mathbf{v}$ and $\mathbf{i}$, this appears to be a coincidence. The important result is that the predictability based on potential predictable components underestimates the predictability of the regression forecast distribution by a factor of 2. This result contradicts our hypothesis that only a small number of potential predictable components can capture the full predictability. The example given here, however, is more appropriately compared with short time weather forecasts, rather than with climate forecasts. We suspect that our hypothesis is valid for large dimensional climate systems on long time scales.

8. Summary

This paper proposed a predictability theory framework that accounts for imperfect forecast models. The critical quantity in this framework is neither the perfect model distribution, which is unknown anyway, nor the accessible forecast distribution, whose state space and variability may differ from the true system, but rather the conditional distribution of the state given all accessible forecasts. This idea also was proposed in Schneider and Griffies (1999), although our interpretation appears to differ from theirs. We have called this distribution the regression forecast distribution because, in the case of normal distributions, it is equivalent to a linear regression of the verification given the accessible forecast. Theoretically, the regression forecast distribution is not the best possible prediction. The best prediction is the conditional distribution given all forecasts and all antecedent observations. However, this latter distribution is independent of the accessible forecasts, reflecting the fact that an imperfect forecast is irrelevant if a perfect forecast model is available. The usefulness of imperfect forecasts appears to lie in the fact that they capture nonlinear or nonstationary behavior, which are difficult to capture in low-dimensional, statistical models estimated from short historical records.

This paper showed that the average predictability of the regression forecast distribution, denoted $I(\mathbf{V}; \mathbf{F})$, satisfies certain fundamental inequalities. First, $I(\mathbf{V}; \mathbf{F})$ provides a rigorous lower bound to the average predictability of the true system. This bound clarifies an important role of accessible forecasts in predictability studies. Second, $I(\mathbf{V}; \mathbf{F})$ is bounded above by the average potential predictability of the forecast model, defined as the average predictability of the accessible forecast distribution relative to its own climatology. The potential predictability of an accessible forecast system does not constitute an upper bound on the true predictability, as is sometimes asserted in the literature. In fact, the potential predictability and the true predictability need not have any relation to each other. Rather, the potential predictability constitutes an upper limit to the average predictability of the regression forecast distribution.

The absence of perfect models has lead some authors to suggest that all measures of predictability require some reference to accessible forecast models; that is, a true predictability does not exist. But defining predictability with respect to forecast models leads to multiple definitions of predictability, one for each model. Also, one should be careful not to equate existence with accessibility: just because we do not have access to something does not mean it does not exist. The framework proposed here presumes the existence of a true predictability, which is the distribution that a perfect model would produce given the initial condition distribution (which itself is constructed from a data assimilation procedure using the perfect model). The true predictability is an inaccessible property of the climate system and associated observations. Accessible forecasts provide lower bound estimates of true predictability; they are not needed to define predictability. The framework correctly implies that classical, deterministic models are perfectly predictable if both the initial condition and dynamical model are known perfectly, but not otherwise. The framework also accounts for the model dependence of uncertainty in the initial condition. True predictability can never be quantified definitively since

at any given time only imperfect forecasts and finite observations exist, and there is no plausible way to eliminate the possibility that a better forecast model or better observations could lead to greater predictability.

This paper suggested that mutual information between accessible forecast and verification $I(\mathbf{V}; \mathbf{F})$ provides an attractive measure of forecast skill. This measure arises naturally in our framework as the average predictability of the regression forecast distribution. It also measures the degree of dependence between two sets of variables that is more fundamentally related to predictability than mean square error, which requires, for instance, that the forecast and verification be represented in the same state space. Furthermore, this measure is invariant with respect to nonlinear transformations of the data, is invariant with respect to the role of forecast and verification, cannot be improved by manipulating the forecast (provided the manipulation is independent of verification), and vanishes if and only if the forecast is statistically independent of the verification. In the case of bivariate normal distributions, mutual information reduces to familiar measures of skill based on the correlation between forecast and verification.

This paper showed that, under certain plausible assumptions, potential predictable components, and these components alone, completely describe the variability of regression forecasts. Potential predictable components therefore provide a basis for reducing the dimensionality of the predictability problem without loss of generality, provided they can be identified. If the forecast and observations are joint normally distributed, then potential predictable components can be obtained by canonical correlation analysis. In non-Gaussian cases, CCA may still provide a useful method of finding predictable components since it optimizes the correlation coefficient.

The predictability of regression forecast distributions for stationary, Gaussian, Markov systems was examined. The distribution of all relevant random variables was given explicitly. If ensemble forecasts are available, the regression forecast distribution varies only with the sample ensemble mean forecast, while the ensemble spread influences the predictability of the regression forecast distribution.

Simple numerical experiments were conducted to illustrate the above concepts and to gain insight into the usefulness of regression forecast distributions. In these experiments, the truth was identified with a single realization from a chosen model, while the forecast was generated by a model from the same class but with different parameter values. The exact predictability of a regression forecast distribution of a two-dimensional, Gaussian, Markov model was computed for various ensemble sizes. The results revealed that relatively small increases in predictability were to be gained with increasing ensemble size. This conclusion is not surprising since, for joint normal distributions, the regression forecast distribution depends on the forecast ensemble only through the ensemble mean, and hence the "extra" ensemble members merely "sharpen" the estimate of the ensemble mean. Sample estimates of the predictability of these stochastic systems, derived from numerical realizations, were biased upward relative to their true values. This bias, which is a manifestation of artificial skill that occurs in statistical prediction, represents a significant problem in the estimation of predictability from finite samples. In most cases, the true predictability estimated from finite realizations tended to be larger than the predictability of regression forecast distributions, even for large ensemble sizes. Further investigation of linear stochastic models in different parameter regimes suggest that, in the absence of prior information, imperfect forecast systems are not always useful in joint normally distributed systems since greater predictability often can be obtained directly from data. By contrast, the Lorenz (1963) model revealed opposite behavior: Gaussian approximated regression forecasts generally have more predictability than Gaussian approximated perfect model distributions. This difference was attributed to the nonlinear dynamics in the Lorenz model, which could be captured by an accessible forecast model from the same model class but with slightly incorrect parameter values, but not by a joint normal distribution, which essentially assumes a linear relation between verification and initial condition. Since the historical record is not adequate for developing anything more than simple linear regression laws, we concluded that the present usefulness of imperfect forecast models lies in the extent to which they capture relevant nonlinear dynamics and/or nonstationary behavior, or facilitate the identification of potential predictable components. These experiments on low-dimensional systems did not support the hypothesis that a truncated set of potential predictable components can capture most the predictability. Whether this finding holds in more realistic, large-dimensional climate models on long time scales can only be settled by experiment.

The problem of estimating the predictability of systems from finite samples will be addressed in more detail in Part III.

Acknowledgments. I am very much indebted to J. Shukla, Tapio Schneider, Michael Tippett, and Ben Kirtman for instructive discussions. Comments from

Tapio Schneider, acting as reviewer, and two other reviewers also led to improvements in this paper. Discussions with Lenny Smith also led to helpful clarifications. This research was supported by the NSF (ATM9814295), NOAA (NA96-GP0056), and NASA (NAG5-8202).

REFERENCES

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

DelSole, T., 2004a: Predictability and information theory. Part I: Measure of predictability. *J. Atmos. Sci.,* **61,** 2425–2440.

——, 2004b: Stochastic models of quasigeostrophic turbulence. *Surv. Geophys,* **25,** 107–149.

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

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

Joe, H., 1989: Relative entropy measures of multivariate dependence. *J. Amer. Stat. Assoc.,* **84,** 157–164.

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.

Kloeden, P. E., and E. Platen, 1999: *Numerical Solution of Stochastic Differential Equations*. Springer, 636 pp.

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.

——, 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.,* **26,** 636–646.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

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

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, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.