Comparison of Information-Based Measures of Forecast Uncertainty in Ensemble ENSO Prediction

YOUMIN TANG
Environmental Science and Engineering, University of Northern British Columbia, Prince George, British Columbia, Canada

RICHARD KLEEMAN
Courant Institute of Mathematical Sciences, New York University, New York, New York

ANDREW M. MOORE
Ocean Sciences Department, University of California, Santa Cruz, Santa Cruz, California

(Manuscript received 10 October 2006, in final form 8 May 2007)

ABSTRACT

In this study, ensemble predictions of the El Niño–Southern Oscillation (ENSO) were conducted for the period 1981–98 using two hybrid coupled models. Several recently proposed information-based measures of predictability, including relative entropy ($R$), predictive information (PI), predictive power (PP), and mutual information (MI), were explored in terms of their ability of estimating a priori the predictive skill of the ENSO ensemble predictions. The emphasis was put on examining the relationship between the measures of predictability that do not use observations, and the model prediction skills of correlation and root-mean-square error (RMSE) that make use of observations. The relationship identified here offers a practical means of estimating the potential predictability and the confidence level of an individual prediction.

It was found that the MI is a good indicator of overall skill. When it is large, the prediction system has high prediction skill, whereas small MI often corresponds to a low prediction skill. This suggests that MI is a good indicator of the actual skill of the models. The $R$ and PI have a nearly identical average (over all predictions) as should be the case in theory.

Comparing the different information-based measures reveals that $R$ is a better predictor of prediction skill than PI and PP, especially when correlation-based metrics are used to evaluate model skill. A “triangular relationship” emerges between $R$ and the model skill, namely, that when $R$ is large, the prediction is likely to be reliable, whereas when $R$ is small the prediction skill is quite variable. A small $R$ is often accompanied by relatively weak ENSO variability. The possible reasons why $R$ is superior to PI and PP as a measure of ENSO predictability will also be discussed.

1. Introduction

A crucial component of any prediction system is the ability to estimate the predictive skill of a forecast so that the uncertainty of an individual forecast can be quantitatively estimated practically. This can be achieved by ensemble prediction; that is, repeating a prediction many times, each time perturbing the initial conditions of a forecast model. Through ensemble prediction, the shape of the prediction probability density function (PDF) that describes the prediction uncertainty associated with initial uncertainty can be estimated.

The earliest work using ensemble ideas to explore numerical weather prediction (NWP) uncertainty are documented by Epstein (1969) and Gleeson (1970). Over the last three decades or so, there has been intensive study in NWP and climate predictability using ensemble prediction. These studies can be roughly categorized into ensemble construction, ensemble representation, and ensemble verification. There are at present
two typical methods used to optimally perturb the initial conditions for constructing ensemble forecasts: breeding vectors and singular vectors (e.g., Toth and Kalnay 1993; Molteni et al. 1996). Another practical ensemble method is to randomly take an anomaly from previous analyses of the prediction model itself or other models [e.g., National Centers for Environmental Prediction (NCEP) reanalysis] to perturb the initial conditions (e.g., Derome et al. 2005; Kharin and Zwiers 2001).

The central issue in ensemble representation and verifications is to seek a predictor of forecast skill, by which the degree of confidence that can be placed in an individual forecast can be assessed. The technique widely used has primarily been the first or second moment of ensemble prediction, a measure of ensemble mean or ensemble spread of the prediction PDF (e.g., Buizza and Palmer 1998; Whitaker and Loughe 1998; Moore and Kleeman 1998; Scherrer et al. 2004; Tang et al. 2007a). However, methods using either the ensemble mean or ensemble spread as the measure of forecast uncertainty have often met with challenges and limitations. For example, Tang et al. (2007a) and Kumar et al. (2000) found that ensemble spread is not a good predictor of forecast skill for ensemble climate predictions. On the other hand, a good relationship between ensemble spread and prediction skill was found in some NWP and climate models (e.g., Buizza and Palmer 1998; Whitaker and Loughe 1998; Moore and Kleeman 1998; Scherrer et al. 2004).

Recently, a new theoretical framework for quantifying potential predictability has been developed and applied to examine ENSO and seasonal climate predictability (Schneider and Griffies 1999; Kleeman 2002, 2008; Tippett et al. 2004; Tang et al. 2005; DelSole 2004, 2005), using an approach built on the information theory (Cover and Thomas 1991). In a perfect model scenario, the potential predictability is a good indicator of the actual skill of the model. For any dynamical prediction there are always uncertainties in the specification of initial conditions and boundary conditions. The influence of these uncertainties on prediction skill may be described by two PDFs: the prediction distribution $p$ and the climatological or equilibrium distribution $q$. The potential predictability of a model may be assessed in terms of differences between $p$ and $q$. Such differences can be expressed in terms of (i) absolute entropy (called predictive information by Schneider and Griffies 1999 and DelSole 2004); (ii) a prediction information-deduced quantity (called predictive power by Schneider and Griffies 1999); (iii) relative entropy (called prediction utility by Kleeman 2002); or (iv) mutual information (Cover and Thomas 1991; DelSole 2004). DelSole (2004) discussed mathematically the relationship between these metrics and evaluated their ability to quantitatively measure the predictability in a perfect model scenario. However, it is not clear which measures of predictability and potential predictability are good indicators of actual skill.

In this paper, we will explore the aforementioned information-based measures of predictability using two realistic ENSO models. Our goal is to try to answer the question posed above from a practical view. The emphasis of this paper will be on a comparison of each information-based measure of predictability in terms of the capability to quantify potential prediction utility. This paper is structured as follows: section 2 briefly describes the models, and ensemble generation method used. Section 3 defines the metrics of prediction skill. In section 4, the information-based measures are introduced and interpreted, and include relative entropy, predictive information, predictive power, and mutual information. The central issue as to which information-based metric is a better predictor of forecast skill will be explored in sections 5 and 6. A summary and discussion are given in section 7.

2. Prediction models and ensemble generation methods

Two hybrid coupled models (HCMs), an Ocean General Circulation Model (OGCM) coupled to a statistical atmosphere (called HCM1 hereafter), and the same ocean model coupled to a dynamical atmospheric model of intermediate complexity (call HCM2 hereafter), were used for this study. The different atmospheric components in the two HCMs allow us to examine a theoretical framework to measure the ENSO predictability in more general terms, and to confirm the robustness of results across model formulations.

The ocean model used is Ocean Parallelise (OPA) version 8.1 (Madec et al. 1998), configured for the tropical Pacific Ocean between 30°N–30°S and 120°E–75°W, with horizontal resolution 1° in the zonal direction and 0.5° within 5° of the equator, smoothly increasing to 2.0° at 30°N and 30°S in the meridional direction. There were 25 vertical levels with 17 concentrated in the top 250 m of the ocean. The time step of integration was 1.5 h and all boundaries were closed, with no slip conditions.

The statistical atmospheric model is a linear model, which predicts the contemporaneous surface wind stress anomalies from sea surface temperature anomalies (SSTA); whereas the dynamical atmospheric model
consists of a Gill-type process-based model which computes global anomalies relative to the observed seasonal cycle of surface wind and mean atmospheric wind at 850 mb.

In both coupled models, the OGCM was forced by the sum of the associated wind anomalies computed by the atmospheric model and the observed monthly mean climatological winds. Full details of each model are given in Moore et al. (2006). ENSO prediction skill and predictability in the two HCMs has been documented in Tang et al. (2005).

The stochastic optimals (SOs; Farrell and Ioannou 1993; Kleeman and Moore 1997) were used to construct ensembles of ENSO predictions for the two HCMs. The SOs represent uncertainties associated with stochastic events in the coupled ocean–atmosphere system that can be amplified by the dynamical model during the forecast interval, which in turn leads to forecast error growth.

For white noise in time, the stochastic optimals are the eigenvectors of the operator \( S = \int_0^\tau A^\tau(t, 0)UA(0, t) \, dt \). (1)

Here \( \tau \) is the forecast interval of interest, assumed to be 12 months in this study, \( A(0, t) \) is the forward tangent linearized model of the system from time 0 to time \( t \), \( A^\tau(t, 0) \) is the adjoint of \( A(0, t) \), and the matrix \( U \) defines a norm of interest. In this study, we use a seminorm defined as the square of the Niño-3 SSTA.

In the present study, the SOs were taken to represent uncertainties associated with stochastic events in the coupled ocean–atmosphere system that can be amplified by the dynamical model during the forecast interval \( \tau \), which in turn leads to forecast error growth. A detailed description of the SOs used in the present study can be found in Moore et al. (2006) and Tang et al. (2005). Ensembles were generated by forcing each ensemble member by a randomly chosen combination of the SOs scaled to have an amplitude consistent with observed interseasonal variability in the tropical Pacific. This method was used because of the growing body of literature suggesting that a significant fraction of ENSO variability is stochastically forced by intraseasonal variability.

The ensemble size for the two HCMs was 31 including a control run. The initial conditions for each ensemble member were derived from a 3D variational assimilation system developed by Tang et al. (2003) that assimilated NCEP reanalysis subsurface temperature. With data assimilation both HCMs exhibit a predictive skill that is comparable to some of the best ENSO prediction models for lead times of 12 months (Tang et al. 2003).

3. Metrics of prediction skill

Typical metrics used to evaluate prediction skill include the correlation coefficient, root-mean-square error (RMSE) and some of their derivatives, as defined below.

1) The correlation skill \( r \) is traditionally defined as

\[
    r(t) = \frac{\sum_{i=1}^{N} [T_p^o(t) - \mu^o_T][T_i^o(t) - \mu^o_T]}{\sqrt{\sum_{i=1}^{N} [T_p^o(t) - \mu^o_T]^2} \sqrt{\sum_{i=1}^{N} [T_i^o(t) - \mu^o_T]^2}},
\]

where \( T \) is the index of interest, \( p \) denotes a prediction, \( o \) denotes observations, \( t \) the lead time of the prediction, \( \mu^p \) the mean of the predictions, \( \mu^o \) the mean of observations, and \( N \) the number of samples used.

The contribution of each prediction to \( r(t) \), denoted as \( C \), can be measured by

\[
    C_i(t) = \frac{1}{r(t)} \frac{\sum_{i=1}^{N} [T_p^o(t) - \mu^o_T][T_i^o(t) - \mu^o_T]}{\sqrt{\sum_{i=1}^{N} [T_p^o(t) - \mu^o_T]^2} \sqrt{\sum_{i=1}^{N} [T_i^o(t) - \mu^o_T]^2}} \times 100\%.
\]

2) RMSE is used to measure the error of an individual prediction for all leads up to 12 months, and is defined as

\[
    \text{RMSE} = \sqrt{\frac{1}{12} \sum_{i=1}^{12} [T_i^o(t) - T^o(t)]^2}.
\]

3) Correlation of individual predictions (CIP) is an individual prediction correlated with its observed counterpart during the prediction period of 12 months; that is,

\[
    \text{CIP} = \frac{\sum_{i=1}^{12} (T_i^p - \mu^p_T)(T_i^o - \mu^o_T)}{\sqrt{\sum_{i=1}^{12} (T_i^p - \mu^p_T)^2} \sqrt{\sum_{i=1}^{12} (T_i^o - \mu^o_T)^2}}.
\]
4. Information-based measures of predictability

Suppose that the Niño-3 SSTA index is modeled as a random variable $T$ whose climatological or equilibrium PDF is $q(T)$. In many practical situations there is considerable knowledge of the climatological PDF from long-term historical observations. In a perfect model scenario, we assume that the “observed” state can be any one of the ensemble members predicted by the models. The climatological PDF is then obtained from the model forecasts over all periods. The ensemble prediction for a given initial condition produces a forecast PDF, denoted $p(T)$. The extent to which the forecast and climatological distributions differ is an indication of potential predictability. There is, of course, no predictability when the forecast and climatological distributions are identical. Several useful measures of the difference between $q(T)$ and $p(T)$ from information theory are the relative entropy $R$ or Kullback–Leibler distance, prediction information, predictive power, and mutual information (Cover and Thomas 1991; Kleeman 2002; Schneider and Griffies 1999; DelSole 2004), as defined below.

\[ R = \frac{1}{2} \ln \left( \frac{\det(\sigma^2_p)}{\det(\sigma^2_q)} \right) + \text{tr} \left[ (\sigma^2_p)^{-1} (\sigma^2_q)^{-1} \right] + (\mu^q - \mu^p)^\top (\sigma^2_q)^{-1} (\mu^q - \mu^p) - n \right). \]

Here, $\sigma_q$ and $\sigma_p$ are the climatological and predictive covariance matrices, respectively, while $\mu^q$ and $\mu^p$ are the climatological and predictive mean state vectors of the system, and $n$ is the number of degree of freedom. From (7), we can deduce that $R$ is composed of two components: (i) a reduction in climatological uncertainty by the prediction [the first two terms plus the last term on the right-hand side (rhs) of (7)] and (ii) a difference in the predictive and climatological means [the third term on the rhs of (7)]. These components can be interpreted respectively as the dispersion and signal components of the utility of a prediction (Kleeman 2002). A large value of $R$ indicates that more information that is different from the climatological distribution is being supplied by the prediction, which could be interpreted as making it more reliable.

b. Absolute entropy, predictive information, and predictive power

An alternative measure of the predictability is defined by the entropy difference between the posterior (e.g., prediction) and prior (e.g., climatology) distributions, that is, the difference of absolute entropy between two distributions (DelSole 2004; Schneider and Griffies 1999). Using the same notation as above, the predictive information is defined as

\[ PI = -\int q(x) \ln[q(x)] \, dx + \int p(x) \ln[p(x)] \, dx. \]

The first term on the rhs of (8) denotes the absolute entropy of the prior distribution, measuring the uncertainty of a prior time when no extra information is provided from observation or model; whereas the second term represents the absolute entropy of the posterior distribution, measuring the uncertainty after the observation and associated prediction becomes available. Thus a large PI indicates that the posterior uncertainty will decrease because of useful information being provided by a prediction, that is, the prediction is likely to be reliable.

Schneider and Griffies (1999) further defined predictive power (PP) based on PI, namely,

\[ PP = 1 - \exp(-PI). \]

PP has the same interpretation as PI, namely that a smaller PP corresponds with a more uncertain prediction, and vice versa.

For a Gaussian distribution, PI and PP are readily simplified (DelSole 2004), as below:

\[ PI = \frac{1}{2} \ln \left( \frac{\det(\sigma^2_q)}{\det(\sigma^2_p)} \right). \]
are scalar variances. In this case, $R$, PI, and PP can be simplified as follows (Kleeman 2002; Schneider and Griffies 1999; DelSole 2004):

$$\text{PP} = 1 - \left( \frac{\det \sigma_p^2}{\det \sigma_q^2} \right)^{1/2}. \quad (11)$$

For a univariate state vector with a climatological mean of zero, the covariance matrices in (10) and (11) are scalar variances. In this case, $\mu$, PI, and PP can be estimated from a 20-yr run of the model, leading to similar results.

$$\text{PI} = \frac{1}{2} \ln \left( \frac{\sigma_p^2}{\sigma_q^2} \right), \quad (12)$$

$$R = \frac{1}{2} \ln \left( \frac{\sigma_p^2}{\sigma_q^2} \right) + \frac{\sigma_p^2}{\sigma_q^2} + (\mu^p)^T (\sigma_q^{-1})(\mu^p) - 1$$

$$= \frac{1}{2} \ln \left( \frac{\sigma_p^2}{\sigma_q^2} \right) + \frac{E(x^2)}{\sigma_q^2} = \text{PI} + \frac{E(x^2)}{\sigma_q^2}. \quad (13)$$

$$\text{PP} = 1 - \left( \frac{\sigma_p^2}{\sigma_q^2} \right)^{1/2}. \quad (14)$$

Comparing (12) with (13) and (14) reveals that $R$ contains the information contained in PI. A distinguishing difference between $R$ and PI (and PP) is that $R$ also involves the ensemble mean whereas PI and PP are determined only by the ensemble prediction spread and climatological spread. PI and PP can be understood as a generalization of the univariate error-ratio predictability measures of multivariate states with arbitrary probability distribution (Schneider and Griffies 1999). They are similar to traditional predictability measures defined by the ratio of the RMSE of a prediction to the RMSE of the climatological mean prediction (Murphy 1988; Stern and Miyakoda 1995). Alternatively, $R$ can be interpreted as the average signal-to-noise ratio.

Another notion that was proposed to measure average predictability by DelSole (2004) is Mutual Information (MI), which is defined as the average of relative entropy $R$ or predictive information PI, over all initial conditions. It has been proven mathematically that $R$ and PI have the same average equal to MI (DelSole 2004). Unlike $R$, PP and PI which measure the predictability for a single forecast, MI measures the average predictability of all predictions.

In this study, we focus on the prediction of the Niño-3 SSTA index. Thus, (12), (13), and (14) will be used directly to calculate each information-based measure. The climatological mean ($\mu_q$) and variance ($\sigma_q^2$) were estimated from all ensemble members and years (sample size is $72 \times 31 \times 12$) as in Tippett et al. (2004).$	ext{1}$

---

1 Alternatively the climatological mean and variance were also estimated from a 20-yr run of the model, leading to similar results.
semble mean, which is absent in PP and PI. The large magnitude of the ensemble mean in strong ENSO events also yields a large $R_2$; $R_2$ decreases monotonically with time, with the largest values at a lead time of 1 month and smallest values at a lead time of 12 months. Monotonicity is an important property of predictability measures since it characterizes the nature of predictability that prediction model utility always declines (monotonically) with the length of the forecast due to chaos. Theoretically the monotonicity only holds for $R$ and not absolute entropy for a stationary, Markov process (Kleeman 2002; DelSole 2004). However, for a normally distributed, stationary, Markov process, PI and PP are also monotonic. In Fig. 1, PI and PP are monotonic but in Fig. 2 such monotonic variations are not always held for PI and PP, suggesting that the ENSO system in HCM1 can be better described as a normally distributed, stationary, Markov process than that in HCM2.

Another important property of $R$ and PI is that both have the same average overall initial conditions, and this average is precisely equal to MI. MI measures the average predictability of a system, and decreases monotonically with time. As shown in Fig. 3, this property holds reasonably well for the two models, especially for HCM1. The reason why the two curves do not exactly agree in Fig. 3 is mainly because of the fact that: (i) the ensembles of the HCMs are not exactly Gaussian, which is required for (12)–(14) to hold; and (ii) the curves are obtained using a finite sample size, leading to errors of the estimates.

Figure 4 depicts the scatterplots of MI versus model skill measured in terms of $r$ and RMSE. Figure 4 indicates that MI is a good predictor of model skill, especially for HCM1 where model skill is correlated well with MI. When MI is large, the skill is invariably good (high correlation and low RMSE) whereas when MI is small, the skill is usually low.

For Gaussian variables, MI and $r$ are theoretically related according to DelSole (2004):

$$MI = \frac{1}{2} \ln (1 - r^2). \quad (15)$$

Figure 5 compares the MI from the average of PI and the MI estimated from (15). The difference between the two estimates indicates how well the Gaussian approximation works in the two HCMs. As can be seen, the two estimates of MI are in good agreement with each other, especially for HCM1. Their correlations are 0.98 and 0.75 for HCM1 and HCM2, respectively. Thus (15) is a good indicator of measuring overall skill.

Shown in Fig. 6 are several MIs (i.e., average $R$) obtained using different ensemble sizes. Figure 6 shows that MI is relatively insensitive to the ensemble member, and that as few as 20–30 members may yield meaningful MI. The practical significance of this conclusion relates to the possible use of MI to estimate model skill without running a large number of forecasts. As discussed in Fig. 4, a good relationship is suggested between model skill and MI. This is especially useful for
evaluating fully coupled general circulation models that are very costly to run.

6. The relationship between potential prediction utility and prediction skill

Figures 7a and 8a show the RMSE of each ensemble mean prediction for HCM1 and HCM2, and Figs. 7b–d and 8b–d display the variation of average $R$, $PP$, and $PI$ over all lead times of 12 months accordingly\(^2\) (denoted as $\bar{R}$, $\bar{PP}$, and $\bar{PI}$ hereafter). A comparison of RMSE and $\bar{R}$ reveals that a large relative entropy is often associated with good prediction skill (i.e., small RMSE), especially for HCM1; whereas when relative entropy is small, the skill is much more variable. This is very similar to a so-called triangular relationship that was found to characterize the relationship between ensemble spread and ensemble mean skill in NWP and climate prediction (e.g., Buizza and Palmer 1998; Xue et al. 1997; Moore and Kleeman 1998), namely, that when the ensemble spread is small, the skill is invariably good, whereas when spread is large, the skill is much more variable. Therefore, we will also use the “triangular relationship” to describe the relationship between

\(^2\) Note that the average is with respect to lead time, rather than with respect to the initial conditions that produce MI.
relative entropy and the RMSE skill. Such a “triangular relationship” is evident in a scatterplot of $\overline{R}$ versus RMSE, as shown in Figs. 9d and 10d.

The scatterplot of correlation of individual predictions, CIP, with $\overline{R}$ is shown in Figs. 9a and 10a, where CIP was calculated using (5). As can be seen, the “triangular relationship” is most clear when the correlation-based metrics is used to quantify model predictability. When relative entropy is large, the CIP is generally large meaning that the prediction is skillful whereas when it is small, the prediction is likely to be uncertain. It should be noted that the limited samples used in (5) may be a concern, but the consistent results from both CIP and RMSE can place more confidence in the “triangular relationship” between relative entropy and model predictability. On the other hand, it might be more problematic if we added samples by increasing the lead time of prediction, since there is little skill of prediction for lead time greater than 12 months.

An interesting issue is the fact that some small $R$s can also lead to good skill such as small RMSE. This seems counterintuitive to the notion of relative entropy since a small value of relative entropy suggests little extra information will be provided by the prediction. To explore this, we examined all predictions with RMSE < 0.5 and with $\overline{R} < 0.5$. It was found that almost all of these cases have relatively weak ENSO anomalies in the observations and predictions, leading to small RMSEs. On the other hand, a weak ENSO anomaly suggests that it is close to climatology, leading to a small relative entropy by definition.

There is no obvious discernible relation between PI or PP and model skill in Figs. 8 and 9. For HCM1, a “triangular relationship” similar to that between $\overline{R}$ and model skill might be able to characterize $\overline{P}$ versus...
model skill, as shown in Figs. 9c,f. However for HCM2 (Figs. 10c,f), there is no such a similar “triangular relationship.” Interestingly PI and PP show an opposite “triangular relationship” to RMSE skill in HCM2 as shown in Figs. 10e,f, namely, that a small predictive information or predictive power often corresponds with a small RMSE; whereas when they are large, the RMSE skill is much more variable. Obviously this is a striking counterexample to the notion of prediction information and predictive power, suggesting that the two measures might not be as good a predictor of model predictability as relative entropy in our cases.

We have explored the relationship between the information-based measures and model predictability, and found that there is a discernible “triangular relationship” between relative entropy and model predictability, and that predictive information and predictive power are not as good as relative entropy for quantifying the predictability. Next we further explore these measures using more quantitative analysis methods.

Figure 11 is the scatterplot for HCM1 and HCM2 showing, respectively, the variations of the contribution C of each ensemble prediction to R, PP, and PI for all leading times of 1–6 months. A striking feature shown in Fig. 11 is that most prediction samples are distributed in an area with very small C, indicating that most predictions account for a very small C. The area of small C is usually coincident with small R but has a broad range.
of variation of PI and PP. Similarly, a large $R$ generally corresponds to a large $C$ but a few small $R$ can also lead to large $C$. Thus Fig. 11 exhibits the aforementioned “triangular relationship.” It might be argued that a linear relationship could also be drawn between $R$ and $C$ from Fig. 11 (upper panel) because of the fact that in most cases, small (large) $C$ corresponds to small (large) $R$. It is true, as supported by Fig. 12, showing significant correlation coefficients between $R$ and $C$ for both HCM1 and HCM2. Therefore, the “triangular relationship” between $R$ and $C$ contains a major component of linear relationship. In contrast to relative entropy, the predictive information and predictive power show little relationship to $C$ as shown in Fig. 11 and Fig. 12.

To further verify the “triangular relationship” between $R$ and prediction skill, we explored two kinds of prediction skill using bootstrap tests, classified by a threshold value of $R$, denoted by $R_0$. The first kind of skill is evaluated using all prediction samples with the $R > R_0$, whereas the second is obtained using all prediction samples with the $R < R_0$. Thus for a bootstrap test given a $R_0$, a new subset of prediction samples was obtained. A 1000-member ensemble correlation was computed using the subset samples. Each member of correlation was obtained using 90% of samples that were randomly selected from the subset. When the 90% changes to 80%, the bootstrap test is little changed.
five bootstrap tests for the both cases respectively with different $R_0$. For the first kind of prediction skill, the $R_0$ is respectively the median value of all prediction samples ($R_{0m}$), 1.1$R_{0m}$, 1.2$R_{0m}$, 1.3$R_{0m}$, and 1.4$R_{0m}$; whereas for the second kind of skill, the $R_0$ is $R_{0m}$, 0.9$R_{0m}$, 0.8$R_{0m}$, 0.7$R_{0m}$, and 0.6$R_{0m}$. Shown in Fig. 13 are the averaged correlation skills of the lead times of the first 6 months from these bootstrap tests. As can be seen, when prediction samples with $R < R_0$ are used, the computed correlation skill is small with a relatively
large extent of uncertainty. The smaller the $R_0$, the more variable the computed correlation skill. However, when prediction samples with $R > R_0$ are used, the correlation skill is large with little uncertainty. Thus the bootstrap tests validate the “triangular relationship” discussed above.

Figures 14 and 15 compare the correlation skills of Niño-3 SSTA between two groups of predictions classified by median value of $\mathbf{X}$, $\mathbf{PI}$, or $\mathbf{PP}$, respectively, for both HCM1 and HCM2. The first group has prediction samples with the measure greater than its median value (dashed–dotted line), and the second with the measure
less than the median value (dotted line). For comparison, the original correlation skill calculated using all samples is also displayed (thick line). The median value was chosen so as to yield a nearly equal sample size in either group. As can be seen, it is apparent that the prediction skill with large relative entropy is significantly larger than that with small relative entropy. The vertical bar in Figs. 14 and 15 is an estimate of the...
correlation error bar by the bootstrap method, which measures the extent of the uncertainty in the computed correlation due to the change in sample size. The results show that an increase in the correlation skill results from the contribution of more skillful predictions with larger $R$, rather than from the uncertainty of the finite sample size.

An interesting feature in Figs. 14a and 15a is the asymmetric structure of the correlation. When the predictions with $R > R_0$ are removed, the model skill degrades very significantly; whereas for a nearly equal number of predictions but with $R < R_0$ removed, the model skill increases relatively little, indicating that the predictions with large $R$ have much more significant impact on skill than those of small $R$, which is consistent with the “triangular relationship” between $R$ and model skill.

Comparing the skills of the two groups in Figs. 14b–c and 15b–c reveals that the model predictability is not clearly characterized by PP or PI. As can be seen, the correlation skills of the two groups that were classified by either PP or PI are only slightly different in both models, especially in HCM2 where the skill of the two groups almost overlaps with the original skill computed from all samples. This suggests that PP and PI are less effective measures of model skill.

Figures similar to Figs. 14 and 15 but using RMSE instead of correlation were also obtained (not shown). The results show that the differences of RMSE between the two groups are not as significant as those of correlation, suggesting that $R$ seems less effective in quantifying the forecast uncertainty when the uncertainty is measured by RMSE. This perhaps indicates the presence of a systematic error in the coupled-model Niño-3 index. This is typical of intermediate and hybrid coupled models with a simple atmospheric component (e.g., Moore and Kleeman 1998). Given the simplistic and economic nature of such models, we are usually content with good predictions of the phase and trends in the Niño-3 index, rather than expecting to capture precisely all the details of both amplitude and phase. On the other hand, a systematic model error could be reduced considerably through some statistical schemes.

---

4 A 1000-member ensemble correlation was computed. Each correlation was obtained using randomly chosen sample pairs of predicted and observed Niño-3 SSTA indices with the same sample size as that used to calculate correlation of the two groups. The standard deviation of ensemble correlation was used to represent the extent of the uncertainty.
Feddersen et al. (1999). Thus we focus mainly on correlation-based measures to evaluate model predictability in this study, as in other similar works (e.g., Moore and Kleeman 1998; Tang et al. 2005).

In summary, the relative entropy \( R \) is a more effective indicator of the uncertainty of ENSO prediction than PI and PP, and a “triangular relationship” is suggested between \( R \) and the model skill. When \( R \) is large, the prediction is typically good whereas when \( R \) is small, the prediction skill is much more variable.

7. Discussion and summary

An important task of ENSO predictability studies is to measure the uncertainty of the prediction, and to determine dominant factors that affect the prediction accuracy. In this paper, several recently proposed information-based measures of predictability were compared in terms of their ability to delineate the relationship with the prediction skill of ENSO, through ensemble prediction using two hybrid coupled models. When the metrics of correlation and RMSE are widely used to evaluate the model prediction skill, they make use of observations. On the other hand, however, these information-based metrics defined by the difference between the prediction and climatological distribution do not make use of observations, thereby measuring the potential predictability. In a perfect model scenario, the potential predictability is a good indicator of the actual skill of the model. The relationship between measures of predictability and actual model skill identified in this study offers a practical means of estimating the potential predictability and the confidence level of actual ENSO prediction.

It was found that the mutual information, defined as the average of relative entropy or predictive information, can measure reasonably well the overall skill. When mutual information is large, the prediction sys-

Fig. 13. The variation scope of correlation skill corresponding to different \( R_0 \)s, estimated using the bootstrap method. The asterisks denote the correlation skills estimated from predictions with \( R \) less than \( R_0 \), whereas the circles denote the skill from those with \( R \) greater than \( R_0 \). For a chosen \( R_0 \), 90% of its samples were randomly extracted to calculate a correlation coefficient, and then this process was repeated 1000 times to obtain 1000 correlation coefficients (see main text).
tem has high predictability and vice versa. The relative entropy and predictive information have the same average over all predictions as expected theoretically.

The relative entropy $R$, predictive information $PI$, and predictive power $PP$ were all found to generally decrease monotonically with lead time in all ensemble predictions. This is consistent with the notion that predictability and model skill usually decrease with increasing lead time, suggesting that all three measures have the potential to be a good indicator of the skill of individual predictions. However, a detailed comparison reveals that $R$ is a more effective predictor of prediction skill than either $PI$ or $PP$. This is especially true when correlation-based metrics are used to measure model skill. $PP$ and $PI$ are very closely related by definition. In general, when $R$ is large, the corresponding ENSO prediction is found to be reliable, whereas when $R$ is small the ENSO prediction skill is found to be much more variable, suggesting that a “triangular relationship” exists between $R$ and model prediction skill. The relationship is less significant between model prediction skill and $PP$ or $PI$.

There are several possible reasons why $R$ is superior to $PI$ and $PP$ in measuring predictability in this study. The first (and the most probable) is that $PI$ and $PP$ relative to ensemble prediction spread and climatological spread are independent of the initial condition in a Gaussian system, whereas the $R$ is not (DelSole 2004). The inherent association of $R$ with initial conditions is of significance for ENSO prediction where the initial condition plays an important role to model skill. The second reason is that $PP$ contains not only the $PI$ (and $PP$) component but the signal component determined by the ensemble mean. In fact, the latter dominates $R$ (Tang et al. 2005), and it has been found that the ensemble mean plays a much important role in determining model correlation skill (Tang et al. 2007a). The contribution of the signal component to predictability is

![Figure 14](image-url)

**Fig. 14.** Correlation skill $r$ for HCM1, calculated using two different groups of samples, classified by the median value of (a) $R$, (b) $PP$, and (c) $PI$. For comparison, the skill calculated using all prediction samples is also shown in thick line. The dashed–dotted line is the skill calculated using the predictions with the measure greater than its median value; whereas the dotted line is the skill calculated using the prediction with the measure smaller than its median value. The vertical bar represents the error bar estimated by a bootstrap method (see main text). $R_0$, $PI_0$, and $PP_0$ are the median values of the corresponding measures and vary with lead time.
absent in PP and PI so that they are less effective in measuring predictability associated with the signal component.

Finally, it may be argued that the ensemble generation method used in this study may be responsible for the difference among the information-based measures. Stochastic optimals were used to construct ensembles of predictions. A stochastic optimal is based on the theory of singular vectors that address the optimal growth of a signal over a time period. Therefore since $R$ is essentially a measure of the signal-to-noise ratio, it is intrinsically superior to PI and PP in measuring the predictability. This is important because there is a growing body of literature suggesting that a significant fraction of ENSO variability is stochastically forced by intraseasonal variability in the tropical Pacific, in which case $R$ is more appropriate than PI and PP. In addition it has been found that $R$ also determines the predictability in other situations. For example, in Tang et al. (2007b), the Arctic Oscillation (AO) predictability was studied using an atmospheric general circulation model and a randomly perturbed ensemble scheme. Their results show that $R$ is still a good predictor of forecast skill, and there exists a similar “triangular relationship” between $R$ and model skill.

The practical significance of the conclusions obtained in this study relates to issuing ENSO probabilistic forecasts operationally. First we can use mutual information to estimate model overall skill without running a large of number hindcast. Second $R$ is superior to PI and PP in measuring the predictability, so the ensemble mean dominating $R$ should play a much more important role than ensemble spread that controls PI and PP. Usually the mean of the distribution can be estimated well with a few samples if the distribution has very small variance in a Gaussian system. Thus, a large ensemble does not seem to be required to generate a prediction and measure the predictability for a dynamical system like ENSO.

Recently proposed and developed information-based measures provide an elegant and powerful framework for quantifying predictability, and have attracted considerable attention because of their particular properties that are absent in traditional measures of predictability. For example, they are invariant with respect to nonsingular linear transformations and monotonic
function with time. DelSole (2004) discussed the connection and relationship of these measures in a theoretical framework but there has not so far been a detailed comparison of these information-based measures in realistic predictions. Using two realistic models and ensemble prediction methods, this work is the first to explore and compare these information-based measures in a realistic framework, which should shed the light on some important issues of predictability.

Acknowledgments. We thank anonymous reviewers for providing insightful comments that lead to many clarifications. This work is supported by the Canadian Foundation for Climate and Atmospheric Sciences (CFCAS) through Grant GR-523 and by the Canada Research Chair program.

REFERENCES


Copyright of Journal of Climate is the property of American Meteorological Society and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.