The IRI was established under a cooperative agreement between NOAA Office of Global Programs and Columbia University. Taiwan joined as the first international partner in 2000.
Proceedings of
The Workshop on Forecast Quality

October 10, 2000
Palisades, New York

International Research Institute for Climate Prediction
*Linking Science to Society*

IRI-CW/01/1
January, 2001

The IRI was established under a cooperative agreement between NOAA Office of Global Programs and Columbia University. Taiwan joined in 2000 as the first international partner of the IRI.

Published by the IRI, 61 Route 9W, Palisades, New York, 10964, USA
Foreword

Forecast quality is a central issue for the IRI and for anyone wanting to use climate forecasts or to invest in their application and development. Commonly asked questions include: How accurate are climate forecasts? Can they really be relied upon? Why are forecasts so often wide of the mark? Is there something better than tercile forecasts?

The IRI needs to have clearly articulated positions on these types of questions, with information that is scientifically sound and communicated in ways that address the concerns of forecast users and other inquirers.

The aims of this internal workshop were to examine the meaning of forecast quality, to review the wealth of expertise in the IRI on the topic, and to develop a clearer, more consistent stance on forecast quality that all staff can understand and represent. The workshop was billed as an important event and most of the IRI scientific staff was able to participate.

The program comprised both presentations and open discussion/planning sessions. Presenters were asked to prepare a short abstract, and these, together with a record of the day's discussions, have been collated into this internal workshop report.

The workshop was a lively start of a more explicit emphasis on the subject of forecast quality within the IRI. It is expected to lead to the development of agendas and task groups to develop the IRI's activities in the field – e.g. on forecast quality research, quality monitoring and reporting, background and web-based information, standard statements for particular forecast products, and specific materials for communicating to user groups.

Reid Basher, Neil Ward
Workshop Organizers
IRI Workshop on Forecast Quality

A workshop to review the IRI’s knowledge base and development needs in respect to seasonal-to-interannual climate forecast quality

Monell Auditorium
9.00 am – 5.15 pm, October 10, 2000

Program

9.00 – 9.30 The Meaning of Forecast Quality, its Importance to the IRI, and the Purposes of the Workshop. Reid Basher

9.30 – 10.00 The Scientific Basis for Predictability and The Forecast Tools That Are Used (dynamical and statistical) Neil Ward, SZ, YT

10.00 – 10.45 Definition of Technical Terms in Forecast Skill and Examples of Forecast Skill Scores. Simon Mason, TB

10.45 – 11.15 Break

11.15 – 11.45 Model Performance in Past Years (Including identification of areas of the tropics where skill is good and of the many other areas where skill is only modest). Tony Barnston, NW, LG, SM

11.45 – 12.05 User Needs and Perceptions of Forecast Quality – Illustrations from the Agriculture Sector. Jim Hansen, JP, AA


12.25 – 13.00 Demonstrations of Progress Toward and Prospects for Improved Forecast Quality, Based on Model Performance in Past Years - New Types of Forecast Outputs. Lisa Goddard, TB, NW, SZ

13.00 – 14.00 Lunch

14.00 – 14.30 Interpretation and Evaluation of Net Assessment (Including targeted for some of the regions and seasons where skill is expected). Tony Barnston, SM, LG, SA

14.30 – 15.00 Some Applications Perspectives on Forecast Quality and Value, Shardul Agrawala

15.00 – 15.30 Discussion: Quality Improvement Options in IRI Operational Forecast and Dissemination Systems. Discussion Leaders Tony Barnston, Chet Ropelewski

15.30 – 16.00 Break

16.00 – 16.30 Discussion: Applications Research to Better Define and Meet User Needs for Forecast Quality. Discussion Leaders Kenny Broad, Reid Basher

16.30 – 17.00 Discussion: Prediction and Forecasting Research to Better Meet Needs for Forecast Quality. Discussion Leaders Steve Zebiak, TB, NW

17.00 – 17.15 Concluding Remarks. Antonio Divino Moura
The Meaning of Forecast Quality, its Importance to the IRI, and the Purposes of the Workshop

Reid Basher

A wide range of public views is held about the nature of forecast quality, ranging from unwarranted confidence to unwarranted rejection. Some imagine that any climate anomalies are predictable, while others believe that the available prediction skill has been badly oversold. Forecasting is difficult, and past events have revealed serious shortcomings in prediction capability - such as with the unexpectedly recurring 91-93 episode, and the deficiencies in predicting the development and decline of the 97-98 event. One peer-reviewed article has argued that the dynamical models, upon which great hope is placed for the future, did not perform any better than traditional statistical forecasts.

To a large extent the debate hinges on how one defines the term “quality.” This can have different meanings to the forecaster and the user. Dynamical seasonal models have scientifically demonstrated skill; but this may not represent useful quality in the hands of a forecast user. Moreover, if a statistical forecast produces a better result, then that is what the user should have.

An example of the problem we face was shown by the most recent published forecasts for East Africa, where the maps of the predicted terciles produced by the IRI and the Nairobi-based Drought Monitoring Centre were markedly different. Even if both were scientifically justifiable, the user can only be left somewhat confused or skeptical about the putative quality of the forecast.

Within the IRI itself, there are also different perspectives on the issue, and it was primarily for this reason that the workshop was initiated. The question “How good are they?” inevitably crops up in some form or another in all IRI activities — whenever model results are being discussed, competing forecasts are being compared, or forecasts are being proposed as a basis for user action or research. Forecast quality characteristics form an integral part of any forecast or forecast application, so it is essential for the IRI to identify quality as a central issue in all its research and capacity building activities. It is especially important that there is close consistency between the IRI’s driving belief in forecast value and the scientific understanding of forecast quality. Further, the IRI can play a leading role in clarifying and communicating the quality issue.
In this talk, the word quality is used as the most general term for describing how good a forecast is or could be. The idea has much wider compass than the usual concern with forecast skill measures. A hierarchy of characteristics of forecast quality can be identified, as follows:

- **Quality of approach**: The quality of the science and methods used – the best physics, the best data assimilation, the best numerical methods, etc.

- **Quality of results**: The quality of the predicted fields, as demonstrated by internal consistency tests and comparisons with observations (this has been the traditional focus of forecast quality study).

- **Quality of process**: The quality of the operational forecasting processes, in terms of ISO 9000 service quality standards or of similar practices of documentation, checking of outputs and process management.

- **Quality of products**: The quality of the forecast products, in terms of both content, accuracy, reliability, regularity, accessibility, presentation, etc.

- **Quality to users**: Quality as perceived by users/stakeholders, judged by them on how the products and services address their needs.

- **Quality in application**: Quality in achieving the sought-after application objective (e.g. gain in yield, reduction of risk), as determined by an appropriate accepted method.

The IRI is a natural home for the study of quality. To make substantial advances, the IRI must grasp the issue strongly — we must live it and breathe it, we must research it, and we must demonstrate it. Our concern for quality must show. Our span of interest necessarily must be wide, and must cover both quality of science and quality of relevance. In the present workshop, we can start the process by sharing the extensive knowledge of the topic that exists in the IRI and getting us all onto a similar plane of understanding. We can then identify the areas where we need to do more work. Potential areas include the development of an R&D agenda on the topic, improvements in operational procedure, joint user and forecaster projects on quality improvement, and upgraded communication on quality issues to our constituencies, e.g. via the website and other publications.
The Scientific Basis for Predictability and the Forecast Tools that are Used

Neil Ward

Anyone planning to promote, support or use climate forecasts will usually be more effective if they are confident that the underpinning science is sound, and are able themselves to feel comfortable in expressing the source of their confidence to others. This lecture considered the essential science that needs to be communicated to arrive at that confidence. Once the confidence is attained, the forecasts themselves have greater quality. Also, by communicating a sense of how predictability arises, a user is better placed to appreciate how some products are scientifically feasible, while others are not.

To gain the confidence of a potential partner/user will require differing levels of detail, depending on the situation. The lecture presented material in a generic way, pointing to the need to adapt the material according to the targeted community.

- There is a need to convince people why it is possible to forecast climate several months ahead given that we cannot forecast weather beyond 2 weeks. This requires the idea that climate is average weather, and this can be modified if forcings on the atmosphere change. Scientists have identified that ocean surface temperatures change sufficiently from one year to the next such that we should change our expectations of climate patterns from one year to the next.

- One of the major sources of changes in ocean surface temperatures is the El Nino Southern Oscillation. Because we have models that can forecast this phenomenon 6-12 months in advance, and because ocean temperatures tend to change only slowly over several months, we are able to foresee to some extent the expected ocean temperature pattern a few seasons in advance, and this allows us to make predictions about the range of climate patterns that we should expect in the coming seasons.

- We can try and draw schematic diagrams than communicate how a change in ocean temperature in one location leads to changes in the wind patterns and rainfall. The winds across the Indian Ocean / Pacific Ocean associated with East Africa wet and dry seasons is an example that is easy to sketch and communicates a lot of the essential atmospheric dynamics.
Having established such conceptual understanding, it is then easier to communicate how the forecasts are made. We can use physical and dynamical models of varying complexity to project forward in time from current ocean-atmosphere conditions to estimate the expected ocean surface temperatures during the coming seasons, and the expected atmospheric patterns given those prevailing ocean conditions. These models estimate the quantities of all atmospheric properties at every location in the world, typically at a grid-spacing of about 250km. It is also possible to use statistical analysis of historical data to identify the relationship between seasonal rainfall totals and conditions in the climate during the previous months. The use of ocean surface temperature as a predictor can be understood from the discussion of the source of predictability.

The ocean surface temperature pattern does not completely constrain the climate patterns that will occur. The atmosphere still has some freedom. Thus, even if we can foresee the ocean surface temperature pattern perfectly, the atmospheric climate forecast must give the likelihood of different outcomes. Our current understanding indicates that it is theoretically impossible to foresee the exact climate pattern that will prevail. However, for some regions of the world, we can very strongly shift the likelihood of certain climate events from one year to the next, and the IRI net assessment global map forecasts produced regularly since 1997 are already demonstrating this.

The above outlines some of the key scientific concepts that can be communicated to help build confidence in the science of seasonal forecasting, and hence enhance the quality of the forecasts themselves. The exact format of the material could vary enormously, from training materials to web based products. It is a great challenge, and solutions can be expected to vary according to target community.
An assessment of forecast quality should consider all aspects of correspondence between forecasts and observations. The most commonly used technical measures of forecast quality are described, and some simple measures of forecast value are introduced.

The simplest measure of forecast quality is to determine whether or not the forecast was “correct”. If forecast involves a simple statement about an event occurring or not occurring (e.g. It will rain on Friday), the correctness of the forecast can be determined easily. There are four possible outcomes:

<table>
<thead>
<tr>
<th>Event occurs?</th>
<th>Warning given?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit</td>
<td>Yes</td>
</tr>
<tr>
<td>Miss</td>
<td>Yes</td>
</tr>
<tr>
<td>False Alarm</td>
<td>No</td>
</tr>
<tr>
<td>Correct Rejection</td>
<td>No</td>
</tr>
</tbody>
</table>

Hits and correct rejections constitute correct forecasts, misses and false alarms are incorrect forecasts.

If the forecast is expressed as a continuous variable (e.g. There will be 1 inch of rain on Friday), the correctness of the forecast can be determined in the same way by defining an acceptable margin of error.

Given a set of forecasts, it is possible to calculate the number of times that the forecasts were correct. The most commonly used measure is the Heidke score (or hit score), defined as:

\[
\text{Heidke score} = \frac{\text{number of correct forecasts}}{\text{number of forecasts}} \times 100\%
\]

\[
\text{Heidke score} = \frac{\text{number of hits and correct rejections}}{\text{number of forecasts}} \times 100\%
\]
A famous example of the use of the Heidke score is Finley’s tornado forecasts. After issuing 100 tornado warnings, forecast performance was as follows:

<table>
<thead>
<tr>
<th>FORECASTS</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tornado</td>
<td>No tornado</td>
</tr>
<tr>
<td>Tornado</td>
<td>28</td>
</tr>
<tr>
<td>No tornado</td>
<td>72</td>
</tr>
</tbody>
</table>

Total: 100 2703 2803

which gives a Heidke score of 96.6%.

\[
\text{Heidke score} = \frac{28 + 2680}{2803} \times 100\% = 96.6\%
\]

If no tornado warnings are issued, an even higher Heidke score can be achieved.

<table>
<thead>
<tr>
<th>FORECASTS</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tornado</td>
<td>No tornado</td>
</tr>
<tr>
<td>Tornado</td>
<td>0</td>
</tr>
<tr>
<td>No tornado</td>
<td>0</td>
</tr>
</tbody>
</table>

Total: 0 2803 2803

Heidke score = \[
\frac{0 + 2752}{2803} \times 100\% = 98.2\%
\]

The problem with the Heidke score is that a high score is achievable using the simplest of forecast strategies if the event forecast is rare (or extremely common). The score needs to be expressed in terms that compare forecast performance with a simple forecast strategy such as climatology, persistence, or random forecasts. A skill score is used to compare the quality of a forecast strategy with that of a reference strategy, e.g. climatology, persistence, or random guessing. The skill score defines the percentage improvement in accuracy over the reference forecast.

The Heidke (hit) skill score counts how many more times the forecast was correct compared to the reference strategy.

\[
\text{Heidke score} = \frac{\# \text{correct} - \# \text{expected correct}}{\# \text{forecasts} - \# \text{expected correct}} \times 100\%
\]

For Finley’s tornado forecasts, the skill score is:

\[
\text{Heidke score} = \frac{(28 + 2680) - 2752}{2803 - 2752} \times 100\% = -86.3\%
\]
One problem with the Heidke skill score is that the misses and false alarms do not consider how bad the forecast was – miss is as good (or as bad) as a near-miss. The Linear Error in Probability Space (LEPS) score was designed to penalize a bad miss more than a near-miss. Instead of scoring a 1 for a hit or correct rejection, and a 0 otherwise, the following table of weights is used for a three-category system.

<table>
<thead>
<tr>
<th>OBSERVATIONS</th>
<th>FORECASTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Above</td>
</tr>
<tr>
<td>Above-normal</td>
<td>0.89</td>
</tr>
<tr>
<td>Normal</td>
<td>-0.11</td>
</tr>
<tr>
<td>Below-normal</td>
<td>-0.78</td>
</tr>
</tbody>
</table>

These weights are optimally defined to ensure that forecasts of climatology AND perpetual forecasts of one category AND random guessing have an expected score of zero.

If the forecaster is confident that above-normal conditions are unlikely, but thinks that normal and above-normal conditions are equally likely, a higher LEPS score can be achieved by forecasting above-normal. The forecaster may therefore be encouraged to hedge. Because the LEPS score encourages hedging, it is not a “strictly proper score”.

All categorical verification measures are flawed because they do not account for possible near-misses across category boundaries, and do not account for the accuracy of the forecasts within a category. Commonly used measures of forecast quality for continuous data are mean error:

\[
\text{mean error} = \text{average forecast} - \text{average observation}
\]

which is a measure of forecast bias, and mean squared error:

\[
\text{mean squared error} = \frac{\text{total of squared errors}}{\text{number of forecasts}}
\]

which is a measure of forecast accuracy. These measures are important in weather forecast verification, but less so in seasonal forecasting because of systematic model errors are usually corrected. Instead, the correlation coefficient is used. Correlation is a measure of association.
rather than of forecast accuracy – i.e. does rainfall increase if forecast rainfall increases? Forecast errors can be large, even with a perfect correlation, because of unconditional biases.

The verification of probabilistic forecasts is more complicated. As soon as a forecast is expressed probabilistically, all possible outcomes are forecast, and so an individual probabilistic forecast cannot be defined as correct or incorrect. However, the forecaster’s level of confidence can be correct or incorrect. If the forecast confidence is correct, the forecast probabilities are defined as “reliable”. It is therefore of interest to assess whether the forecaster is over- or under-confident?

Forecast reliability is achieved given two conditions:

- A forecast is consistent if the forecast probability is a true estimate of the forecaster’s level of uncertainty.
- If forecasts are reliable, the forecaster’s confidence is appropriate.

If forecasts are consistent and reliable, the probability that the event will occur is the same as the forecast probability.

Reliability and attributes diagrams are used to indicate forecast reliability. For all forecasts of a given confidence, the relative frequency that the event occurs is calculated and plotted. If the proportion of times that the event occurs is the same as the forecast probability, the probabilities are reliable (or well calibrated). Unfortunately, a large number of forecasts are required to construct a reliability diagram.

Because climatological forecasts are reliable (unless the climate is changing), it is desirable for forecast probabilities to be sharp (to differ from the climatological probability) as well as reliable.

Some verification scores provide combined measures of reliability and sharpness. The Brier score (sometimes called the half Brier score) is one such example. It measures the mean-squared error of probability forecasts.

\[
\text{Brier score} = \frac{\text{total of squared probability errors}}{\text{number of forecasts}}
\]

For example, if the event was forecast with a probability of 60%, the probability error is 60% - 100% = -40%. The ranked probability score is similar to the Brier score, but is for multiple categories. The Brier score and the ranked probability score can be expressed as skill scores in the same way as for the Heidke (hit) score.
Verification measures for continuous probabilistic forecasts are experimental – there are very few attempts to estimate the full probability distribution of possible outcomes.

These verification measures are designed to measure the quality of a set of forecasts over a period of time. Can these methods be applied to verify a set of forecasts for a specific period? The first consideration is that there must be a sufficient number of independent forecasts for the score calculations to be meaningful. If there are, non-probabilistic forecast measures do have validity. However, the interpretation of probabilistic verification measures is problematic: if events with low probabilities occur over large areas, verification of the forecast probabilities will suggest, possibly incorrectly, that the probabilities are unreliable.

Forecast quality does not guarantee that forecasts are of value to users. One problem is that the cost of forecast errors may be asymmetric. For example, given a forecast stating that there will be one inch of rain on Friday afternoon, is it better to have 2 inches than no rainfall? A miss may have a different size of impact than a false alarm.

The relative (or receiver) operating characteristics (ROC) are used to estimate whether forecasts are potentially useful. If forecast confidence exceeds a threshold minimum a warning is issued. If (proportionately) more warnings are issued for events than for non-events, then the forecasts are potentially useful. The ROC compares the hit rate (proportion of events for which a warning was provided correctly) to the false-alarm rate (proportion of non-events for which a warning was provided incorrectly). The threshold forecast confidence is varied to calculate a set of hit and false alarm rates that are then plotted on a graph.
Model Performance in Past Years (including identification of areas of the tropics where skill is good and of the many other areas where skill is only modest).

Tony Barnston

Dynamically-based atmospheric general circulation models (AGCMs) are the predominant tools used to make the IRI’s seasonal climate forecasts. A forecast for the global SST is first made, and then the AGCMs are run using the forecast SST fields. In order to know how much confidence to place in the atmospheric climate forecasts made by the AGCMs, it is necessary to know how they would have performed if used to predict the climate over past years, for which the observed outcomes are known. Such simulations have been made for the 1950-95 period for the three models currently used: The ECHAM-3, the CCM-3 and the NCEP-MRF-9. The behavior of sets of 10 ensemble members for each model forecast, and the implied levels of skill of the ensemble mean for selected seasons and regions, were discussed. The 10 ensemble members were run using the same forecast SST, but different atmospheric initial conditions in order to get some idea of the resulting variation of forecast outcome.

It was noted that the simulations were done using simultaneous, observed SST (the so-called AMIP design) – unlike what would be possible in a real-time forecast. To more accurately assess expected model skills, persisted SST, or preferably even model-forecasted SST, would need to be used in the hindcasts. Still, the expected skills for the immediately forthcoming season would not be expected to be severely overestimated in many cases, since SST anomalies often persist for at least several months. With this caveat understood, the following model capabilities were identified.

Model skills are high enough to be useful only in certain seasons and regions, and in most cases there is a known physical underpinning for those skills. For example, skill potential is high over northeast Brazil during the February through May period due to ENSO-related SST and the SST in the tropical Atlantic Ocean. Moderately high skills are shown for eastern Africa, eastern Australia and much of Indonesia during October through December; in the Sahel, Sudan and Gulf of Guinea of Africa in July through September; and in central Chile, southeastern Brazil and Uruguay in November through February. It is more modest, but possibly useful, in other regions/seasons. The ENSO phenomenon is responsible, directly or indirectly, for most, but not all, of the seasonal climate predictability. Predictive skill over Africa has its origin both in ENSO and in more local SST patterns, in which certain of the latter may be related to ENSO to some degree.
The degree of skill ranges from forecast versus observation correlations of over 0.7 in the case of northeastern Brazil in March through May, to the 0.4 to 0.7 range for many of the other regions/seasons mentioned above. These skill levels are high enough to make it worthwhile to issue forecasts, but imperfect enough that the associated uncertainties require use of a probabilistic forecast format. The uncertainty comes about because, despite the consistent effect of the relevant SST anomalies, the occurrence of unpredictable individual weather events within the season represents a random “noise” factor that can greatly influence the seasonal rainfall total or mean temperature. The forecast probabilities express the degree to which a user can count on the category of climate anomaly that is given the highest probability of occurrence. In many cases, even the most likely tercile has less than a 50% probability (but greater than the climatological 33% probability). While such forecasts are expected to have utility over a longer period, their benefit may not be apparent in a single cases or even in a small set of cases such as two or three years for the season/region in question.

Some discussion took place regarding possibilities for forecasting periods shorter than a season and spatial scales smaller than the typical large-scale regions that have homogeneous responses to given SST anomaly patterns. While much research has yet to be done, and the answers to these questions are still open, it appears likely that mainly the statistical aspects of the more finely spatially and temporally resolved events will be able to be forecast. For example, the variability of raininess may be able to be forecast to some extent, but not the temporal phasing of the wet spells.

Further progress in downscaled forecasting, and in the skill of the large-scale seasonal forecasts, is likely. However, the inherent limit of climate predictive skill is not known, and may turn out to be not much higher than what we see presently.
Agricultural production is sensitive to climate, particularly solar radiation, temperature and dynamics of precipitation. Seasonal climate forecasts may allow farmers to reduce risk by modifying management to reduce losses or benefit from opportunities. However, potential benefits depend on farming systems, farmer characteristics and resources, and forecast characteristics beyond meteorological quality. Several forecast characteristics are generally important to farmers. The first is site specificity. Farmers are aware of spatial variability, can recognize scale mismatches between forecasts and decisions, and want to know what to expect on their field(s). Second, concerns about temporal specificity include rainy season onset, dry spells, and harvest conditions. Availability relative to decisions is critical. The third – interpretive information – is controversial. Farmers sometimes solicit help in interpreting management implications of forecasts. In other cases, farmers seek only forecasts, and have clear ideas about using them. Possible determinants may include indigenous forecast tradition, prior research, provider’s capabilities and track record, perceived flexibility, and methodology. The authors’ opinion is that interpretive information founded on appropriate research and participatory evaluation will usually enhance forecast utility. Fourth, the source of information influences usefulness. Farmers can best utilize information from sources they know and trust, who are able to both disseminate forecasts and support their interpretation and use. The fifth characteristic is skill. Farmers consistently ask about forecast accuracy (sometimes in ambiguous terms). They are concerned more with temporal consistency than spatial coherence. Communicating the probabilistic nature of imperfect forecasts is a critical challenge.

In an analytical example, the subjective value to a risk-averse farmer of optimal use of ENSO phases for farm land allocation decreased from +$4.07/ha to -$6.74/ha when El Niño and La Niña events were treated as unbiased but deterministic forecasts. In conclusion, farmers are more concerned with the value of forecasts than with meteorologists’ conception of quality. Value is enhanced when information from known, trusted sources matches decisions (e.g., spatial scale; timing relative to decisions, impacts and system dynamics; with adequate and demonstrated skill; and content targeted to decisions). Farmer understanding, confidence and, sometimes, perceived flexibility to act, apparently grow with exposure.
User Needs and Perceptions of Forecast Quality – Illustrations from the Water Resource Sector

Steve Rayner and Upmanu Lall

The public sector and legal and political institutions have long played a direct role in the management and development of water resources. In this setting, the multiple objectives served by water storage facilities, often translate into an operational strategy that reflects a limited enthusiasm for deviation from normative practices that have evolved to manage long-term risks. Existing reservoir operation rules provide generalized guidelines and annual and monthly operating targets that are followed by the manager. There is often some flexibility in short-term operation in the process of meeting long-term targets. The growing involvement of “stakeholders” in watershed management typically succeeds in negotiated short-term storage management to balance competing interests, while satisfying existing long-term contracts and targets. Thus, significant benefits that may be possible from using climate and other social and environmental information to manage storage over time frames consistent with the reservoir fill and drain cycle are difficult to accomplish in practice, because they may require changes in infrastructure.

Water resource managers have the technical capacity to quantitatively analyze and use uncertain information. Hence, the water resource sector is perceived as a promising target for climate forecasts. However, the evolved technical and management structures require a demonstration that the forecasting system is able to reproduce estimates of historical water supply and demand to a legally acceptable accuracy. They also call for verifiable scenarios of water supply and demand at specific spatial locations in a manner that is consistent across the multiple time scales of system operation from flood control to low flow augmentation for ecological health. Public participation and existing institutional strictures lead to a critical examination of the direct relevance of the information presented, as well as an increase in partisan perspectives that exploit different aspects of the uncertain information. Thus, the forecast quality question is translated into a question as to the relevance and specificity of forecast products, a need for benchmarks of long term performance of the user specific products, and for an identification of the types of climate information that are most relevant for watershed management. We argue that such benchmarks and products do not exist, and that existing climate forecast products might find limited application.

Information on the nature of interannual climate variability information may be useful to explore changes in long term operation and planning that reflect a new appreciation of the
nature of hydroclimatic events that a region can experience, and the possibly cyclical nature of their long term variation. Cooperative regional planning may benefit from the qualitative understanding that specific climate events in other parts of the planet may affect regional conditions for several months. It may be possible to adjust long term operational targets in response to seasonal and longer climate forecasts (that have been translated into specific hydrologic scenarios). Experiments with innovative management options that respond to specific stakeholder interests may result from an exploitation of the window of opportunity provided by the political climate surrounded by a well-publicized forecast. There are opportunities for working within existing operational frameworks and developing technologies for system adjustment that use existing climate forecasts in conjunction with historical flow and climate records, and a history of climate forecasts for the location of interest. An illustration of how one could generate regional hydrologic scenarios by appropriately sub-sampling historical flows, while automatically incorporating a consideration of climate forecast accuracy, and then use these scenarios to adjust reservoir “rule curves” is provided.
Atmospheric general circulation models (GCMs) form the primary basis for IRI's official seasonal climate forecasts. These numerical climate models capture much of the observed behavior of the climate; however, they also have many shortcomings. Some of these have been addressed already and incorporated into our forecast operations, and others are being researched to determine the feasibility of further improvements. The idea is to maximize the potentially useful information provided by the currently available state-of-the-art GCMs. We are most concerned with improving the reliability of our probabilistic forecasts as well as the resolution in space, time, and probabilities.

**Improving reliability of forecasts**

Probabilistic forecasts can be obtained directly from GCMs by running the same model many times, forcing the atmosphere with the same sea surface temperature (SST) boundary conditions and modifying only the initial condition of the atmosphere. It is generally not optimal to use these ensemble predictions uncorrected, due to deficiencies in current generation GCMs. Consistent errors in the shape and spread of the model's ensemble probability distribution can often be corrected by adjusting the signal or variance of the ensemble. Long (30+ years) simulations from the GCM compared to observed climate are used to identify where and when these systematic errors are occurring and how they might be corrected. Similarly, systematic spatial errors may exist in the GCMs, and the best method for correcting these is under investigation.

**Improving resolution of forecasts**

By correcting the ensemble probability distribution from the GCM, the probabilities of the model are assumed more representative of the true uncertainty in the current climate system, because the uncertainty introduced by errors in the model have been minimized. This allows for more direct use of the model's probability distribution, yielding potentially more information than what is provided in the tercile probabilities of IRI's official forecast. Large ensembles, and more and/or better GCMs, that are likely to be incorporated in IRI's forecast operations should lead to better estimates of the GCM probability distributions, and thus should also allow for more resolution of the forecast probabilities.
**Improving/modifying forecast format**

With improved reliability allowing for more resolution in the forecast probabilities we would like to add forecast information that addresses the full distribution of climate outcomes - particularly the tails. The problem with supplying specific information about probabilities at the tails of the distribution is that so few of these events have occurred in our historical record (over a period consistent with the GCM simulations). This makes it difficult to accurately correct for the reliability of the models in such situations.

A supplementary format of the forecasts is planned for the near future that will give a better resolved (more than 3-categories) probability distribution as a plot of the probability distribution function and/or the probability of exceedence. However, this type of information must be accompanied with appropriate error bars, caveats, and explanation.

**Prospects for increasing time resolution within the 3-month season.**

Monthly forecasts, onset of rainy season and other higher time-scale predictions are something users want and are starting to get from other sources (often unidentified), but the feasibility of such predictions still hasn't been proved. In some cases, making statements about specific months relative to the season under forecast is related to a mis-match of the climatological season and the forecast season. In general, we have not found the GCM results for potential predictability at higher time scales very promising. More regionally targeted and quantitative studies of feasibility for these predictions are underway. Any added value in this direction will certainly be regionally and seasonally dependent.

- **Monthly forecasts versus seasonal forecasts**
  In general, there is a loss of simulation skill for individual months relative to a 3-month average. The skill at the monthly timescale is likely not separable from the seasonal signal, except when the forecast season does not coincide with a climatologically meaningful 3-month season (e.g. climatologically no rain in the first or last month of forecast season). The main information for any of the three months, in regions of potential predictability and where the model has skill, is contained in the seasonal average. Specific cases that were examined include: Indonesia (OND); Africa (JFM - southern Africa; JAS - western Africa; OND - eastern Africa); North America (JFM). The ability to identify differences between the months within the season appears to be a very rare situation, and is currently based on limited statistical studies.
• **Characteristics of weather within the climate**
  A study was conducted that questioned whether a GCM could simulate the shift in frequency of extreme daily events within season. The region examined was eastern Australia, a semi-arid climate with a distinct and variable rainy season that is potentially predictable and that is strongly impacted by heavy rainfall events.

  Two models were examined: ECHAM (moderate skill) and NCEP (good skill). The hit rate was just barely statistically significant for predicting increases in extreme daily events in some seasons, even in the better model. The hit rate for decreased event frequency may be significant for some seasons, but this is likely not adding more information to a forecast of Below-Normal for the season.

  Further investigations will be conducted using regional GCMs, statistical methods, and statistical-dynamical hybrid methods to see if it may be possible to add these higher time-scale predictions to the seasonal forecasts.
Interpretation and Evaluation of Net Assessment (including targeted for some of the regions and seasons where skill is expected)

Tony Barnston

The Net Assessment forecasts are expressed probabilistically with respect to the three terciles of the climatological probability distribution, that distribution currently being based on the 30-year period of 1961-90. When no forecast indications are present (either due to a weak SST forcing, a poor signal-to-noise season/region, low model skill, or several of these), climatological probabilities are given — i.e., a 33.3% probability is assigned for each of the three categories. The tercile boundaries were defined purely by rank (i.e. using the wettest 10, driest 10, and middle 10 years) within the 30-year base period. Typical tercile precipitation forecast probabilities range from 15% to 60%, although occasionally tilts of the odds away from climatology as high as 70% and as low as 5% have been assigned.

The IRI seasonal forecasts (Net Assessments) have been issued every 3 months since the season of October to December (abbreviated “OND”) 1997. The precipitation forecasts have been verified from that season through the OND season of the year 1999, for a total of 9 Net Assessments. This verification, to follow, is simple and preliminary. The magnitude of the probability is not considered, and the category having the highest probability is regarded as the forecast category. This interpretation sets a poor example for users on how the probability forecasts should be regarded, since it treats the forecast as if it were a deterministic forecast for only one of the three tercile categories. It is used here for simplicity and because the results are qualitatively very similar to more sophisticated verification measures designed for probability forecasts. Comparing the forecast category to the category of the observed rainfall, either a hit or a miss is defined. The number of hits is then tallied and divided by the number of forecasts, where forecasts are aggregated spatially (area-weighted) within one forecast, and also over time to compute an average score over the 9 seasons. With no skill, a score of 0.333 is expected by chance, and with “perfect” skill a score of 1 would be expected. This proportion of hits is then rescaled linearly to become Heidke skill scores, where the 0.333 hit rate is scaled to be 0 and the perfect score becomes 100.

For example, a two-thirds hit rate would produce a Heidke skill score of 50. The Heidke skill score H is formally expressed as 
\[ H = 100 \times \frac{\text{number correct} - \text{total number}}{\text{number expected} - \text{total number}} \]
where number correct is the number of hits, total number is the number of forecasts, and number expected is the number expected by chance in the absence of any real skill. For 100 forecasts, for example, number expected would be 33.3 for the tercile system. Note that negative scores (down to −50 at worst) are possible. The Heidke skill score was used.
in an earlier verification of the IRI's forecasts (Mason et al. 1999, BAMS, 1853-1873). The table on page 22 lists the Heidke scores for each of the nine 3-month target periods from OND 1997 to OND 1999, followed by summary scores over the 9 forecasts making up the 2.25-year period. Skills are computed only for non-climatological forecasts. The percentage of the region having non-climatological forecasts is reflected in the “% coverage” figure in the last column. Summary skills are weighted by the inter-forecast coverage; i.e. individual forecasts with low coverage are weighted proportionately low.

Because of the shortness of the period, the skills shown in the table cannot be used for robust, definitive conclusions regarding the effectiveness of the IRI’s forecast system, but rather as hints or signs. The global skills are the best overall indicators, since they are aggregated over the largest set of areas, allowing random sampling errors to cancel. These show modest but consistently positive skills for forecasts of the first season, and similar but more variable skill for the longer lead forecasts. Their higher variability is likely related to their lower areal coverage, resulting in less cancellation of sampling errors. Skill is intuitively expected generally to decrease with increasing lead time, largely due to less certainty in the SST forecast used to drive the atmospheric models. That the IRI forecasts did not show this trend may suggest that the areal coverage for the first lead time might have been higher than warranted – i.e. confidence in assigning non-climatological probabilities may have been overestimated. This would result in skills lower than what would be possible if smaller portions of the region had been given non-climatological forecast probabilities.

A reliability analysis (Wilks and Godfrey, 2000: “Proceedings of the 24th Annual Climate Diagnostics and Prediction Workshop”, Palisades, New York) indicated that for some regions the forecast probabilities, while fairly conservative, may still have been too far from the climatological probabilities. The skills in some regions suggest that their percentage of forecast coverage may have been overdone (e.g. Europe, northern Asia, the Sahel). Skills for regions/seasons expected to be predictable *a priori* (denoted by asterisks in the table), due to being ENSO impact areas or being sensitive to SST anomalies not related to ENSO, are generally comparatively high despite some cases of negative skills. That is, the skills in these regions are generally in line with what would be expected on the basis of the model simulations whose skills were discussed in an earlier presentation in this workshop. It should be noted that the period examined here had strong ENSO forcing: a strong El Nino from mid-1997 to early June 1998, and a moderate to strong La Nina from mid-June 1998 through 1999. Skills during ENSO-neutral periods are expected to be lower than those during ENSO-active periods. Verification scores for temperature (not shown here) are expected to be generally higher than those for precipitation, in keeping with historical findings.
The IRI’s forecast system is undergoing improvement and refinement as better ways to combine multi-model ensemble forecasts into a single indication, and more powerful computers, are introduced. A computer upgrade in 2001 is expected to enable horizontal resolution of the models to increase from T42 to T63, and the number of models used is expected to increase as the IRI forms partnerships with other organizations both domestic and international. Forecast applications are also expanding, particularly in the area of specially tailored products – some of which require spatially downscaled regional models. A threefold increase in the frequency of forecast issuance is planned for year 2001. Along with these positive changes, it is hoped that forecast skill will come closer to the inherent limit of predictability, and that opportunities to apply the forecasts in new and different ways will be identified in accordance with user needs.
Table 1. Heidke Skill Scores for Recent IRI Tercile Forecasts

Heidke skill scores for forecasts for the first lead forecast season, except for scores of Globe—Seas2 and Tropics—Seas2 (25S-25N), which are for the second (3-month lead) season. The average skill, and the average percentage of the area given non-climatological forecasts (% coverage), are shown in the far righthand columns. Dashes indicate zero coverage. Bold indicates, for relevant regions, the seasons when relatively high predictive skill is expected a priori. The ENSO state is indicated by “+” for El Nino, “-” for La Nina, “0” for near neutral (note sudden change to La Nina in early June 1998). Broad regions are listed first, then smaller regions.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ENSO State</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Globe</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td>4</td>
<td>15</td>
<td>11</td>
<td>16</td>
<td>7</td>
<td>16</td>
<td>12</td>
<td>47</td>
</tr>
<tr>
<td>Globe-Seas2</td>
<td>15</td>
<td>6</td>
<td>-1</td>
<td>34</td>
<td>27</td>
<td>31</td>
<td>-6</td>
<td>13</td>
<td>13</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Tropics</td>
<td>25</td>
<td>26</td>
<td>16</td>
<td>8</td>
<td>14</td>
<td>9</td>
<td>24</td>
<td>10</td>
<td>12</td>
<td>17</td>
<td>65</td>
</tr>
<tr>
<td>Tropics-Seas2</td>
<td>26</td>
<td>12</td>
<td>0</td>
<td>39</td>
<td>23</td>
<td>37</td>
<td>-1</td>
<td>14</td>
<td>19</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>9</td>
<td>17</td>
<td>-3</td>
<td>11</td>
<td>10</td>
<td>30</td>
<td>-7</td>
<td>3</td>
<td>56</td>
<td>11</td>
<td>56</td>
</tr>
<tr>
<td>Europe</td>
<td>25</td>
<td>-26</td>
<td>-15</td>
<td>-5</td>
<td>14</td>
<td>5</td>
<td>6</td>
<td>-18</td>
<td>33</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td>South America</td>
<td>45</td>
<td>43</td>
<td>22</td>
<td>7</td>
<td>11</td>
<td>15</td>
<td>27</td>
<td>-1</td>
<td>42</td>
<td>26</td>
<td>63</td>
</tr>
<tr>
<td>Africa</td>
<td>6</td>
<td>3</td>
<td>20</td>
<td>-16</td>
<td>7</td>
<td>6</td>
<td>37</td>
<td>13</td>
<td>13</td>
<td>10</td>
<td>57</td>
</tr>
<tr>
<td>Austr/NZearl</td>
<td>-3</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>16</td>
<td>11</td>
<td>-11</td>
<td>8</td>
<td>-4</td>
<td>7</td>
<td>53</td>
</tr>
<tr>
<td>Asia, &gt;37N</td>
<td>-16</td>
<td>0</td>
<td>1</td>
<td>-11</td>
<td>13</td>
<td>-15</td>
<td>-8</td>
<td>36</td>
<td>-20</td>
<td>-4</td>
<td>41</td>
</tr>
<tr>
<td>Asia, &lt;37N</td>
<td>0</td>
<td>10</td>
<td>13</td>
<td>-6</td>
<td>21</td>
<td>27</td>
<td>52</td>
<td>1</td>
<td>-17</td>
<td>12</td>
<td>57</td>
</tr>
<tr>
<td>Indonesia</td>
<td>70</td>
<td>47</td>
<td>61</td>
<td>40</td>
<td>-30</td>
<td>34</td>
<td>14</td>
<td>-14</td>
<td>26</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>E Australia</td>
<td>-1</td>
<td>47</td>
<td>14</td>
<td>20</td>
<td>30</td>
<td>59</td>
<td>-23</td>
<td>22</td>
<td>-21</td>
<td>19</td>
<td>80</td>
</tr>
<tr>
<td>NE Brazil</td>
<td>-13</td>
<td>93</td>
<td>63</td>
<td>--</td>
<td>-50</td>
<td>-26</td>
<td>50</td>
<td>6</td>
<td>46</td>
<td>29</td>
<td>82</td>
</tr>
<tr>
<td>E Africa</td>
<td>93</td>
<td>69</td>
<td>20</td>
<td>-27</td>
<td>50</td>
<td>-22</td>
<td>--</td>
<td>100</td>
<td>-6</td>
<td>41</td>
<td>55</td>
</tr>
<tr>
<td>Sahel</td>
<td>-19</td>
<td>-12</td>
<td>51</td>
<td>-8</td>
<td>-50</td>
<td>--</td>
<td>-50</td>
<td>25</td>
<td>-50</td>
<td>4</td>
<td>55</td>
</tr>
<tr>
<td>Guinea Coast</td>
<td>0</td>
<td>-50</td>
<td>25</td>
<td>-50</td>
<td>-50</td>
<td>40</td>
<td>36</td>
<td>25</td>
<td>0</td>
<td>2</td>
<td>78</td>
</tr>
<tr>
<td>SE Asia</td>
<td>22</td>
<td>77</td>
<td>18</td>
<td>-43</td>
<td>28</td>
<td>58</td>
<td>51</td>
<td>19</td>
<td>-18</td>
<td>23</td>
<td>63</td>
</tr>
<tr>
<td>Central Am</td>
<td>29</td>
<td>51</td>
<td>25</td>
<td>14</td>
<td>100</td>
<td>-50</td>
<td>-17</td>
<td>17</td>
<td>-50</td>
<td>26</td>
<td>65</td>
</tr>
<tr>
<td>SE South Am</td>
<td>100</td>
<td>55</td>
<td>54</td>
<td>11</td>
<td>43</td>
<td>-8</td>
<td>34</td>
<td>--</td>
<td>77</td>
<td>50</td>
<td>78</td>
</tr>
<tr>
<td>United States</td>
<td>-5</td>
<td>20</td>
<td>15</td>
<td>-44</td>
<td>29</td>
<td>47</td>
<td>-35</td>
<td>-17</td>
<td>67</td>
<td>13</td>
<td>61</td>
</tr>
<tr>
<td>Canada</td>
<td>12</td>
<td>11</td>
<td>-8</td>
<td>23</td>
<td>-18</td>
<td>46</td>
<td>14</td>
<td>--</td>
<td>-28</td>
<td>7</td>
<td>44</td>
</tr>
<tr>
<td>Trop Pac Is.</td>
<td>87</td>
<td>-2</td>
<td>-50</td>
<td>100</td>
<td>-10</td>
<td>22</td>
<td>47</td>
<td>33</td>
<td>28</td>
<td>27</td>
<td>46</td>
</tr>
</tbody>
</table>
Some Applications Perspectives on Forecast Quality and Value

Shardul Agrawala

This presentation raised five key issues of relevance to the discussion on forecast quality from the perspective of applications research and/or the end-user community. First, there has been a significant shift in recent years from statistical to dynamical model based seasonal forecasts, largely because the latter explicitly model the physical system. However, dynamical model outputs routinely need to be adjusted to correspond better with historical observational data. Whether such statistical adjustments might compromise the quality of dynamical model forecasts needs to be better analyzed.

Second, it is important to evaluate the quality of those forecasts that reach the hands of the users. In this respect, a more rigorous evaluation of the Regional Climate Outlook Fora forecasts that are widely disseminated to user communities might be more critical than IRI Net Assessments.

Third, the quality of forecasts needs to be evaluated at the scales at which the information is likely to be used. There is currently a big mismatch between the regional scale of several thousand square kilometers at which forecasters evaluate the skill of their products and the farm or district level where most users are likely to use such information.

Fourth, from a user perspective, it might be more critical to assess and communicate forecast value, and not just forecast skill. One major constraint in this regard is that current seasonal forecasts only assign probabilities of the total seasonal precipitation falling in three tercile categories. While “skill” might be assessed based upon whether the most likely tercile is realized in terms of the seasonal total, “value” is often critically dependent on the temporal evolution of rainfall within the season and several such scenarios are plausible within seasonal totals that fall under one tercile. Some current cost-loss analyses have tended to overlook this important issue.

Finally, it is important to note is the “clunkiness” of the state of the art forecast product that only gives tercile probabilities, for seasonal rainfall totals, and that too over very large regions. Greater utilization of forecast information might depend less on how we communicate the quality of current forecasts and more on the extent to which we can increase the resolution of forecast information – spatially, temporally, and in terms of associated probabilities.
The discussion on the quality improvement options in IRI operational forecasts and dissemination systems was initiated by the notion that the users of forecast information are more interested in an improved spatial resolution that can at present be produced by the IRI, albeit with modest skill. Additionally, they are more likely to be affected by the lower and upper ends, respectively, of the at present three-tercile probability forecast product. However, some of the first ever forecasts produced globally were for a probability distribution of either below or above the climatological mean for a particular region. Later on, three categories were proposed, and these were the categories used ever since.

Notwithstanding the three-category probability forecast, very little forecast skill exists for the middle tercile, i.e. the near-normal category. A suggestion was made that, instead of the present three-category forecasts, there may be skill in producing forecasts for five categories. If the increased number of categories turns out to be feasible, it could better resolve the aforementioned issue of producing forecasts for the upper and lower ends of the distribution. The idea of offering the forecast probability distribution on a continuum was also proposed, although not everyone agreed that the users are ready for so much information.

Furthermore, it was suggested that only those categories that would be of use to a user of the product, should be produced. However, the vast number of possible users of forecast information may require an equally large number of preferred forecast categories. Even if particular categories could be identified for certain user applications, it is as yet not clear what the uncertainty of the predicted probabilities of the categories will be. That is, the forecast probabilities themselves may have some error. At present, it should be assumed that the assigned probabilities of the categories are reliable, although it has been shown during the workshop that for some cases, the assigned probabilities are over-confident (too far from the climatological probability). In conclusion, it was suggested that the limitations of the current forecast products should be communicated to the users, and that the IRI should also be flexible by using their larger scale products to try to produce more user orientated products through extrapolation to a smaller scale.
Simon Mason mentioned two major directions within forecast R&D: (i) improving the sharpness and reliability of current forecast products through recalibration and adjusting for over (or under) confidence of dynamical models; and (ii) exploring the production of new forecast related products. With regard to the latter Simon indicated that reducing the length of the forecast period appears difficult at present, as there is a loss in skill in going to monthly forecasts. There does however seem to be some potential in establishing links between seasonal rainfall totals and the occurrence of extreme rainfall events within the seasons in some regions of the world.

Tony Barnston mentioned that it might also be possible to go beyond tercile probabilities and produce products such as probability of exceedence graphs that could give users more specific information. Chet Ropelewski, however, expressed concern that more specific information on the probability distribution function might convey a level of knowledge that is not borne out in current model outputs, and might be misleading to some users. Steve Zebiak noted that improvements in dynamical forecasting capability were important regardless of the degree of current use of climate forecasts. However, a one-size-fits-all forecast product cannot be the optimal product for any single user. There might be potential in targeting particular applications and developing tailored forecast products for such applications using dynamical model outputs. Steve also noted that there is no substitute for having a transparent forecast track record.
Discussion: Prediction and Forecasting Research to Better Meet User Needs for Forecast Quality

Discussion Leader: Steve Zebiak
Rapporteur notes: Cintia Uvo

The discussion started by Steve Zebiak raising several points that caught his attention during previous talks of the day. He noted the very pessimistic scenario about the utility of every product of the forecast. It was explained that we are just beginning an age of use of climate information. Products are still in development. Besides, climate information doesn’t change from one year to another. The evolution is more in terms of generations. The society also needs to learn how to use this information. The improvement of climate prediction and the education of the society will happen in parallel.

IRI produces global and general products that are not optimal for the use of any specific user. The IRI has the role of producer and collector of information. IRI should produce the foundation of information that would be available to all users. This foundation of information provides to the user the possibility of defining or developing the product that is interesting and adapted to their needs. A set of different methodologies could be applied to this foundation of information so that more useful and user specific products could be created.

The physical bases for forecast correction was pointed out by Shardul Agrawala during his talk. A good example of the physical basis of the corrections is the topography correction. This is a very needed kind of correction due to the low resolution of the GCMs that not allow a correct representation of the topography by the models.

There is no substitute for track record. A long time series is the best information one can get. The recognition by the user of its experience on a time record will instill trust in these users on the methods to be applied. It is important in the development of objective methods to put all available information (models and statistics) together.

After the raising of these points, individual comments started. Jennifer Phillips asked what is IRI official position — the use of statistical forecasts or net assessment? Steve Zebiak replied that if one method is better than the other, that's the one that should be used. Chet Ropelewski noted that models will eventually do better than statistical methods, but meanwhile let's use what works best.
Joshua Qian replied that much research is still needed so that we will be able to develop products beyond terciles or to increase model accuracy for small scale. Neil Ward stated that there is a natural improvement of the techniques with time and that it is important to find techniques that merge results from GCM experiments and statistics.

Reid Basher asked how to forecast climatic characteristics as onset of rainy seasons. Steve Zebiak replied that there are still aspects that are challenging to GCMs like the capture of the swings of the monsoon. It is even difficult to understand why the models can not do it. Maybe the problem is on shorter time scales. Jim Hansen noted that one of our missions should be enhancing the society capacity to understand climate fluctuation. Decision makers should also identify their needs. It is important to stimulate the collaboration between application people and forecasters. Steve Zebiak emphasized that it is clear that as institution we should collaborate more. For example, we should work together to define new products for specific areas. Evolution forces dialogue in some way, and some application products could be developed this way. Antonio Divino Moura also stated that the IRI should work together in this way. Nowadays we are more successful forecasting climate in some areas than in others. It is important to develop better products to serve applications targeting areas that we already have good results, (e.g. the Ceará Project). All groups should be acting together to solve problems.

Tony Barnston asked Jim Hansen about how we might improve internal collaboration. Jim Hansen said that while the projects he works with do not have much involvement of other groups at the IRI, such as the project in southern India which has no climatologists from IRI, he would like to see more projects where different groups of IRI are committed to developing projects and products together.

Jennifer Phillips noted that we tend to conceive of ourselves as a climate prediction institution and have climate variability forecasts as the goal of the institution. We should develop more products of interest to applications. Which group should produce things that the application group thinks is useful, such as statistics for a certain location? Chet Ropelewski responded that this kind of product could be developed by the monitoring group in conjunction with the modeling group. The problem is to establish the priorities of the products, because there is a whole universe of products that can be done. Reid Basher concluded that what we produce depends on the responses we get from users.
Concluding Remarks

Antonio Divino Moura

Dr. Moura said the workshop had been a good start on mapping out and discussing the issues, and that as we get more clarification and mutual understanding we should bring the matter to a wider audience. Forecasts and their quality are a key component of the IRI construct, and the consideration of quality helped make the connection between modeling and prediction on the one hand and users and applications on the other. He said that the report of the meeting would be an important output of the meeting. He expressed his satisfaction with the day’s talks and discussion and thanked the organizers, speakers, and participants for their contributions.