Science and Technology Infusion Climate Bulletin
NOAA's National Weather Service
36th NOAA Annual Climate Diagnostics and Prediction Workshop
Fort Worth, TX, 3-6 October 2011

Skill of Real-time Seasonal ENSO Model Predictions during 2002-2011 — Is Our Capability Increasing?

Anthony G. Barnston [1], Michael K. Tippett [1], Michelle L. L'Heureux [2], Shuhua Li [1], and David G. Dewitt [1]

[1] *International Research Institute for Climate and Society, The Earth Institute of Columbia University, Palisades, New York*

[2] *Climate Prediction Center, NCEP/NWS/NOAA, Camp Springs, Maryland*

1. Introduction

In this study, real-time model predictions of ENSO conditions during the 2002-2011 period are evaluated and compared to skill levels documented in studies of the 1990s. ENSO conditions are represented by the Niño3.4 SST index in the east-central tropical Pacific. The skills of 20 prediction models (12 dynamical, 8 statistical), that have been displayed on the ENSO prediction plume of the International Research Institute for Climate and Society (IRI) since 2002, are examined. Over the last two to three decades, our ability to predict ENSO variations at short and intermediate lead times has presumably gradually improved due to improved observing and analysis/assimilation systems, improved physical parameterizations, higher spatial resolution, and better understanding of the tropical oceanic and atmospheric processes underlying the ENSO phenomenon. Studies in the 1990s showed moderate ENSO prediction capability, with forecast versus observation correlations of about 0.6 for 6-month lead predictions for the Niño3.4 region (Barnston *et al.* 1994). This study reviews the recent model performances, and reexamines the question of the relative performance of dynamical and statistical models. We also compare the skills of the 9 years of real-time predictions to those of longer-term (30-year) hindcasts from some of the same models. The ENSO prediction models studied here are listed in Table 1.

The ENSO predictions issued each month from February 2002 through January 2011 are examined for multiple lead times for future 3-month target (*i.e.*, predicted) periods. Figure 1 shows the variability of the Niño3.4 anomaly from 1981 to 2011, highlighting the recent 9-year period of the current study. Although there were some moderate ENSO events, no very strong events occurred. The last target period is January-March 2011, while the earliest target period is February-April 2002 for the shortest lead time and October-December 2002 for the longest lead time. The forecast data from a given model consist of a succession of running 3-month mean SST anomalies with respect to the climatological means for the respective predicted periods, averaged over the Niño3.4 region.

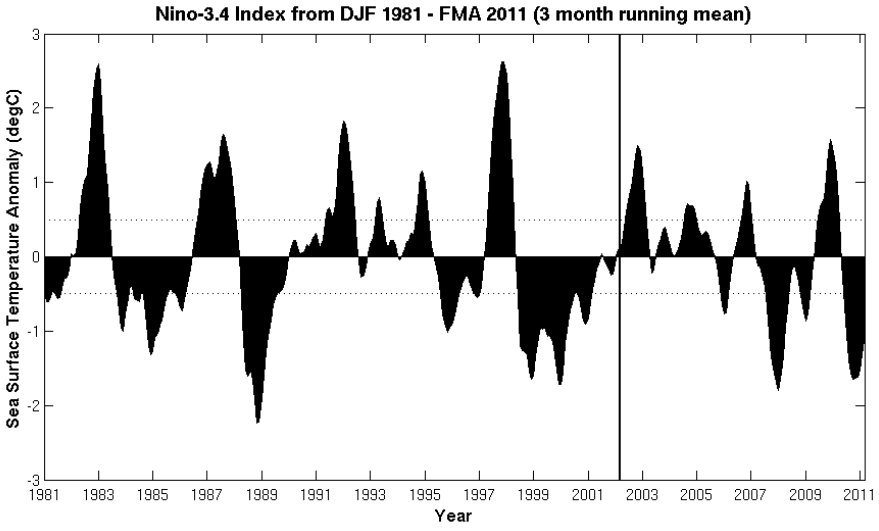

Figure 1 Time series of running 3-month mean SST anomaly with respect to the 1981-2010 period climatology in the Niño3.4 region for 1981-2011, highlighting the 2002-2011 study period.

Correspondence to: Anthony G. Barnston, International Research Institute for Climate and Society, The Earth Institute of Columbia University, Palisades, New York; E-mail: Tonyb@iri.columbia.edu.

Predicted periods begin with the 3-month period beginning immediately after the latest available observed data, and continue for increasing lead times until the longest lead time provided by the given model, to a maximum of 9 running 3-month periods. Here, lead time is defined by the number of months of separation between the latest available observed data and the beginning of the 3-month forecast target period. Although anomalies were requested to be with respect to the 1971-2000 climatology, some prediction anomalies were with respect to means of other periods, such as from 1982 to the early 2000s for some dynamical predictions. Adjustments for these discrepancies were not conducted, nor were model bias corrections attempted. Only the ensemble mean of the dynamical model forecasts is considered as a deterministic prediction.

Dynamical Models	**Model type**
NASA GMAO	Fully coupled
NCEP CFS	Fully coupled
Japan Meteorological Agency	Fully coupled
Scripps Hybrid Coupled Model (HCM)	Comprehensive ocean, statistical atmosphere
Lamont-Doherty	Intermediate coupled
Australia POAMA	Fully coupled
ECMWF	Fully coupled
UKMO	Fully coupled
Korea Met. Agency SNU	Intermediate coupled
Univ. Maryland ESSIC	Intermediate coupled
IRI ECHAM/MOM	Fully coupled, anomaly coupled
COLA Anomaly	Anomaly coupled
COLA CCSM3 (too short a record)	Fully coupled
Météo France (too short a record)	Fully coupled
Japan Frontier FRCGC (short record)	Fully coupled
Statistical Models	**Method and predictors**
NOAA/NCEP/CPC Markov	Markov: Preferred persistence and transitions in SST and sea level height fields
NOAA/ESRL Linear Inverse Model (LIM)	Refined POP: Preferred persistence and transitions within SST field
NOAA/NCEP/CPC Constructed Analogue (CA)	Analogue-construction of current global SSTs
NOAA/NCEP/CPC Canonical Correlation Analysis (CCA)	Uses SLP, tropical Pacific SST and sub-surface temperature (not used beginning in 2010)
NOAA/AOML CLIPER	Multiple regression from tropical Pacific SSTs
Univ. British Columbia Neural Network (NN)	Uses sea level pressure and Pacific SST
Florida State Univ. Multiple Regression	Uses tropical Pacific SST, heat content, winds
UCLA TDC Multi-level Regression	Uses 60N-30S Pacific SST field

Table 1 Dynamical and statistical models whose forecasts for Niño3.4 SST anomaly are included in this study. Note that some models were introduced during the course of the study period, or replaced a predecessor model.

The Reynolds-Smith version 2 optimal interpolation (OI) observed SST data averaged over the Niño3.4 region (5°N-5°S, 120°-170°W) is used as the verification data, using the 1981-2010 period to define the anomalies.

2. Results

a. Real-time Predictive Skill of Individual Models

Time series of the running 3-month mean observed SST anomalies in the Niño3.4 region and the corresponding predictions by 23 prediction models at 0-, 2-, 4- and 6 month lead times are shown in Fig. 2, showing that the models generally predicted the variations of ENSO with considerable skill at short lead times, and decreasing skill levels with increasing lead times. Figure 3 shows the temporal correlation between model predictions and the corresponding observations as a function of target season and lead time, with a separate panel for each model. The correlation skill patterns of the models appear roughly comparable. All indicate a northern spring predictability barrier (Jin *et al.* 2008), with short lead prediction skills having a relative minimum for northern summer, extending to later seasons at longer lead times. Relative to the statistical models, Fig. 3 shows higher correlation skills by many of the dynamical models for seasons in the middle of the calendar year that generally have lowest skill. By contrast, for seasons having highest skills (*e.g.* northern winter target seasons at short to moderate lead times), skill differences among models and between model types appear small.

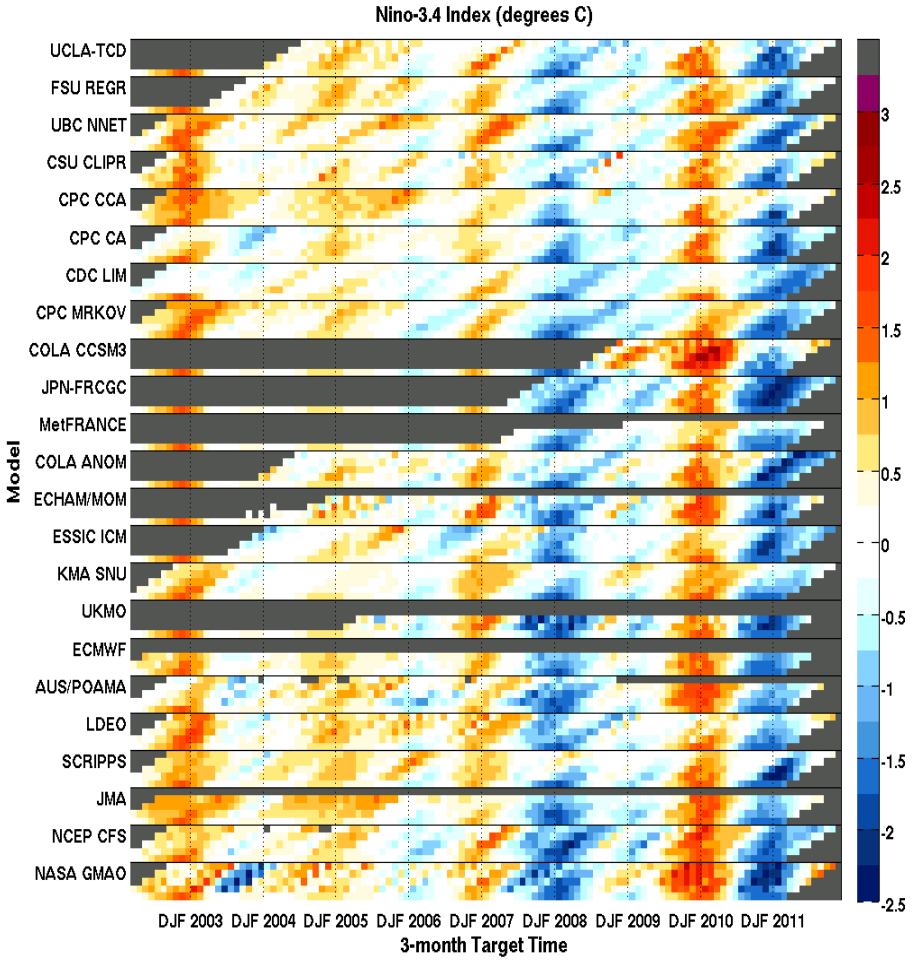

Figure 2 Time series of running 3-month mean Niño3.4 SST observations (°C anomaly), and corresponding model predictions for the same 3-month period from earlier start times at 0-, 2-, 4- and 6-month leads. Data for each model are separated by thin black horizontal lines. The first 8 models at the top are statistical models. For each model, the bottom row shows the observations, and the four rows above that row show predictions at the four increasing lead times. Vertical dotted lines demarcate calendar year changes, separating Nov-Dec-Jan from Dec-Jan-Feb. Observations span from Feb-Mar-Apr 2002 to Jan-Feb-Mar 2011, while forecasts at longer lead times start and end with later seasons. Black shading indicates missing data.

Figure 4 shows individual model correlation skills as a function of lead time for all seasons combined, while the top and bottom panels of Fig. 5 show skills for the pooled target seasons of NDJ[1] , DJF and JFM, and for MJJ, JJA and JAS, respectively. Overall, model correlation skills at 6-month lead range anywhere

[1] Seasons are named using the first letter of the three constituent months; e.g. DJF refers to Dec-Jan-Feb.

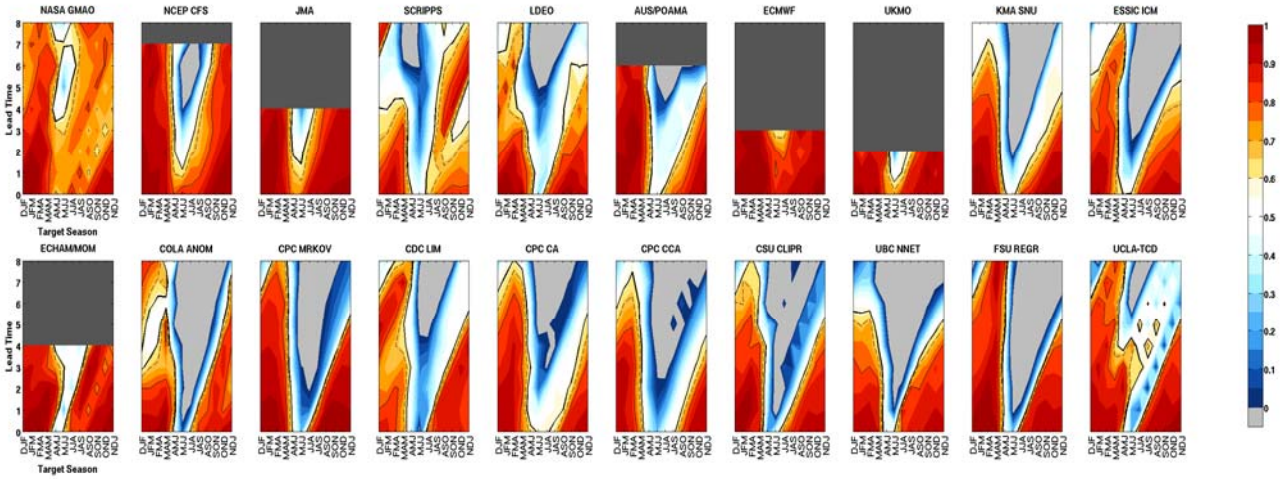

Figure 3 Temporal correlation between model forecasts and observations as a function of target season (horizontal axis) and lead time (vertical axis). Each panel highlights one model. The first 12 models are dynamical, followed by 8 statistical models. Thick solid contour shows the 90% significance level, dashed contour the 95% level, and thin solid contour the 99% level

from zero to about 0.7 for all seasons combined, while predictions for the northern winter season range from 0.4 to 0.9, and for the northern summer season from below zero to 0.55. The model skill levels for all seasons combined (Fig. 4) differ from one another noticeably at all lead times. Averaged over all seasons, skills average somewhat lower than the 0.6 level found at 6-month lead in earlier studies. However, a small number of current models, some of which do not predict out to 6 months lead, have shorter-lead skill levels that would exceed a 0.6 correlation if their forecast range were extended, and if their skill followed a downward slope with increasing lead time averaging that shown by other models having longer maximum lead times. Examples of models with such good or potentially good skill include ECMWF, NASA-GMAO, JMA; NCEP-CFS skill approximately equals 0.6. However, two caveats in the comparison of skills of today's models against models of 10 to 20 years ago include (1) the ENSO variability during the 2002-2011 period will be demonstrated to have been more difficult to predict than that over the 1981-2011 in general; and (2) the current set of predictions were made in real-time, while those examined in previous studies were partly hindcasts. Both factors will be examined further below.

One reasonably might ask whether the skill differences at any lead time are sufficient, for a 9-year period, to statistically distinguish among the performance levels of some of the models. Because ENSO episodes last up to a year, we assume (perhaps conservatively) that we have only about one independent sample per year. The existence of statistically significant differences between skills of any pair of individual models requires very large sample skill differences—larger than those found here. However, the statistical significance of skill differences between dynamical and statistical model types is more tractable, and is addressed below.

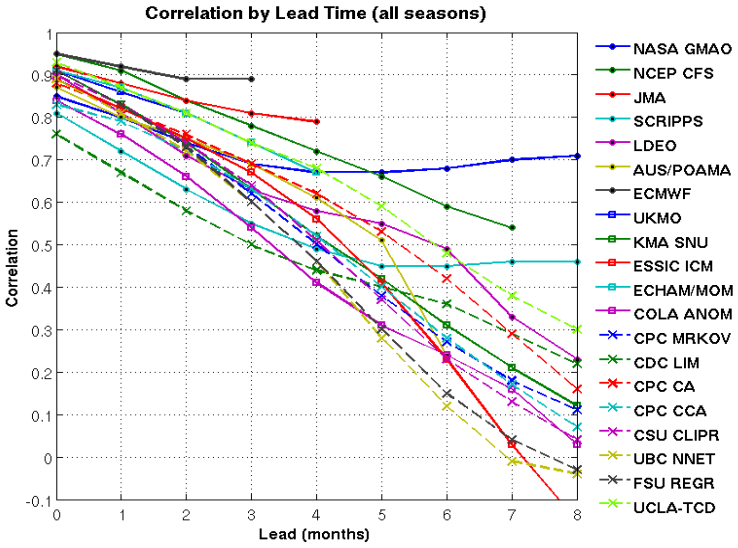

Figure 4 Temporal correlation between model forecasts and observations for all seasons combined, as a function of lead time. Each line highlights one model. The 8 statistical models are shown with dashed lines and the cross symbol.

The correlation between model

predictions and observations reflects purely the discrimination ability of the models, since biases of various types do not affect this metric. However, such prediction biases (*e.g.*, calibration problems involving the mean or the amplitude of the predictions) are also part of overall forecast quality, despite being correctible in many cases. To assess performance in terms of both calibration and discrimination, root mean square error (RMSE) is examined. Here the RMSE is standardized for each season individually. Standardization scales RMSE so that climatology forecasts (zero anomaly) result in the same RMSE-based skill (of zero) for all seasons, and all seasons' RMSE contribute equally to a seasonally combined RMSE. Figure 6 shows RMSE as a function of lead time for all seasons together. The ECMWF model has the lowest RMSE over its range of lead times. For lead times greater than 2 months, persistence forecasts have higher RMSE than that of any of the models. There is clearly some comparability between correlation skill (Fig. 4) and RMSE (Fig. 6), with models having highest correlation tending to have low RMSE. However, exceptions are discernible, due to the effects of mean biases and amplitude biases (not shown).

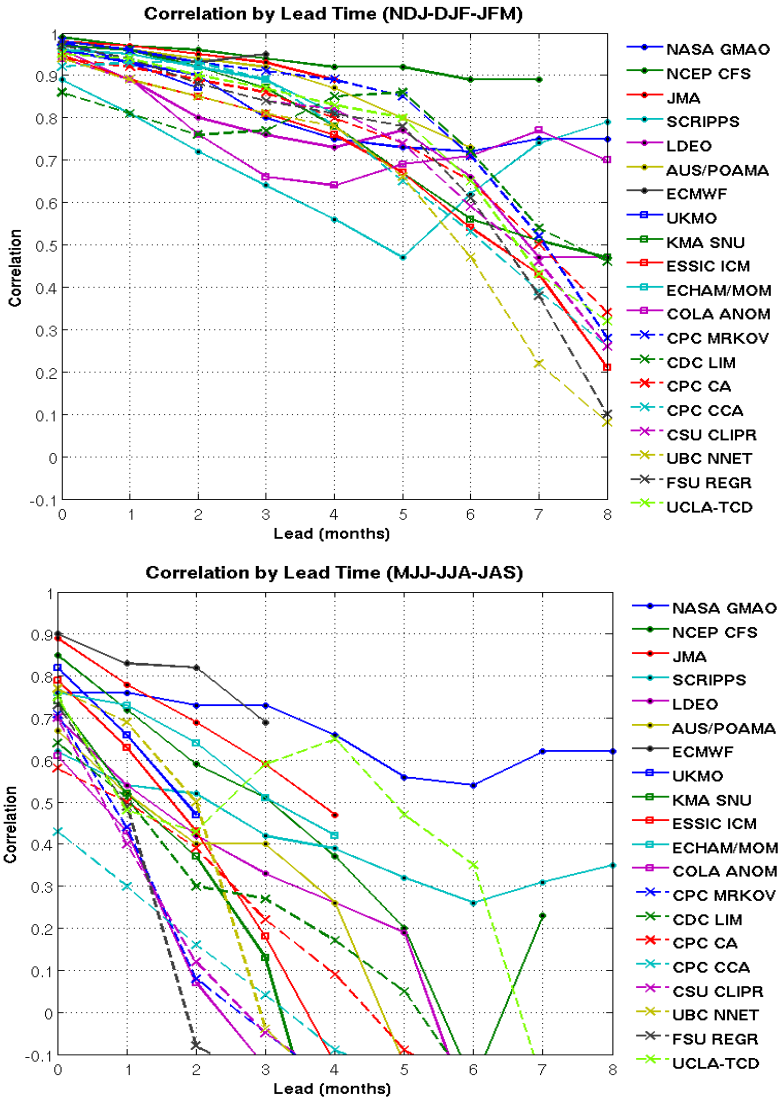

Figure 5 (top) Temporal correlation between model forecasts and observations for Nov-Dec-Jan, Dec-Jan-Feb and Jan-Feb-Mar as a function of lead time. Each line highlights one model. The 8 statistical models are shown with dashed lines and the cross symbol. (bottom) As in top, but for May-Jun-Jul, Jun-Jul-Aug and Jul-Aug-Sep.

Establishing statistical significance of skill differences between dynamical and statistical models for specific times of the year is difficult for a 9-year study period. However, the fairly large number of models can be used to help overcome the short period length. Models are ranked by correlation skill for each season and lead time separately, using the 9 year sample. Systematic differences in the ranks of the dynamical and statistical models are identified using the Wilcoxon rank sum test (Wilcoxon 1945). Additionally, the average correlation of the dynamical and statistical models is compared using a standard t-test, applied to the Fisher Z equivalents of the correlations (Ramseyer 1979). The p-values resulting from these two statistical approaches are shown in Table 2. Although the difference-in-means test generally yields slightly more strongly significant results than the rank sum test, the season/lead patterns of the two approaches are similar. Significant differences, in which dynamical models tend to outperform statistical models, are found at short lead time for the target periods near May-Jul-Jul, the seasons just following (and most strongly affected by) the northern spring predictability barrier. This significance pattern migrates to later target periods with increasing lead time, following the target periods corresponding to the fixed forecast start times of April or May. For forecasts whose lead times do not traverse the northern spring barrier, statistical versus dynamical skill differences are not significant.

Wilcoxon rank sum test (field significance p=0.034)													
Lead	**DJF**	**JFM**	**FMA**	**MAM**	**AMJ**	**MJJ**	**JJA**	**JAS**	**ASO**	**SON**	**OND**	**NDJ**	**All**
0	0.32	0.19	0.76	0.28	**0.01**	**0.01**	0.09	0.95	0.22	0.41	0.95	0.76	0.70
1	0.88	0.25	0.22	0.32	0.06	**0.003**	**0.02**	0.17	-0.68	0.34	0.68	1.00	0.64
2	1.00	0.76	0.32	0.19	0.17	**0.04**	**0.01**	**0.01**	0.54	-0.73	0.38	0.94	0.22
3	1.00	-0.93	0.32	0.14	0.36	0.19	0.14	**0.01**	**0.01**	0.25	1.00	-0.74	0.12
4	-0.33	-0.37	0.79	0.29	0.48	0.48	0.18	0.25	**0.01**	**0.004**	0.36	-0.21	0.16
5	-0.09	-0.29	-0.40	0.67	0.92	0.60	0.67	0.30	0.34	**0.03**	**0.002**	0.92	0.21
6	0.60	-0.60	-0.75	1.00	0.40	0.46	0.75	-0.83	0.75	0.46	**0.05**	**0.05**	0.25
7	**0.02**	1.00	-0.35	1.00	-0.64	0.20	0.82	0.91	0.70	0.73	0.25	**0.03**	0.35
8	**0.05**	**0.02**	1.00	-0.52	-0.44	1.00	0.19	-0.61	-0.66	-0.88	1.00	**0.05**	0.61

t-test for mean difference (field significance p=0.026)													
Lead	**DJF**	**JFM**	**FMA**	**MAM**	**AMJ**	**MJJ**	**JJA**	**JAS**	**ASO**	**SON**	**OND**	**NDJ**	**All**
0	0.27	0.19	0.65	0.22	**0.003**	**0.001**	0.06	0.48	0.41	0.65	0.74	0.46	0.49
1	0.50	0.37	0.12	0.16	**0.04**	**0.000**	**0.01**	0.10	-1.00	0.32	0.73	-0.85	0.29
2	-0.90	0.54	0.33	0.06	0.23	**0.04**	**0.001**	**0.01**	0.16	-0.86	0.29	0.93	0.12
3	-0.98	0.65	0.13	0.13	0.71	0.20	0.07	**0.002**	**0.001**	0.11	-0.90	-0.65	0.09
4	-0.35	-0.39	0.82	0.40	0.26	0.78	0.29	0.12	**0.002**	**0.000**	0.22	-0.19	0.11
5	-0.16	-0.47	-0.31	0.56	-0.93	0.55	0.87	0.17	0.18	**0.004**	**0.001**	0.66	0.18
6	0.34	-0.80	-0.73	-0.62	0.34	0.30	0.70	0.97	0.66	0.28	**0.01**	**0.02**	0.18
7	**0.01**	0.37	-0.39	-0.60	-0.67	0.26	0.42	0.65	0.42	0.51	0.17	**0.01**	0.15
8	**0.04**	**0.02**	0.52	-0.37	-0.45	-0.83	0.13	0.77	0.48	0.66	0.70	0.29	0.34

Table 2 Statistical significance results (2-sided p-values), by target season and lead time, for differences in temporal correlation skill of dynamical versus statistical models: (top) Wilcoxon rank sum test for correlation skills, and (bottom) t-test of difference in means of Fisher Z equivalents of the correlations skills. Entries statistically significant at the 0.05 level are shown in bold. Negative sign indicates cases when statistical models have higher ranks (or means) than dynamical models. P-values are shown to 3 decimal places when p<0.005; 0.000 indicates p-value <0.0005.

Although significant differences are noted for specific seasons and leads, there is a multiplicity of candidate season/lead combinations, and 5% of the 108 candidates (*i.e.*, 5 or 6 of them) are expected to be significant by chance. In the case of the Wilcoxon test, 20 entries are significant, and for the difference-in-means test 20 entries are significant. To assess the field significance of the collective result (Livezey and Chen 1983), Monte Carlo simulations are conducted in which the model type is randomly shuffled 5,000 times, maintaining the actual number of dynamical and statistical models for the given lead time, and the set of local significances is regenerated. Using the sum of the z or t values of all 108 cells as the test statistic, the percentage of the 5,000 randomized cases that exceeds the actual case is determined. The z or t values are

taken as positive when the correlation of the dynamical models exceeds that of the statistical models, and negative for the opposite case. Resulting field significances are 0.034 and 0.026 for the Wilcoxon rank test and t-test, respectively, indicating significantly low probabilities that the set of local significances occurred accidentally. This finding suggests that the circumstance under which local significance is found, namely forecasts impacted by the northern spring predictability barrier being more successful in dynamical than statistical models, is meaningful and deserves fuller explanation.

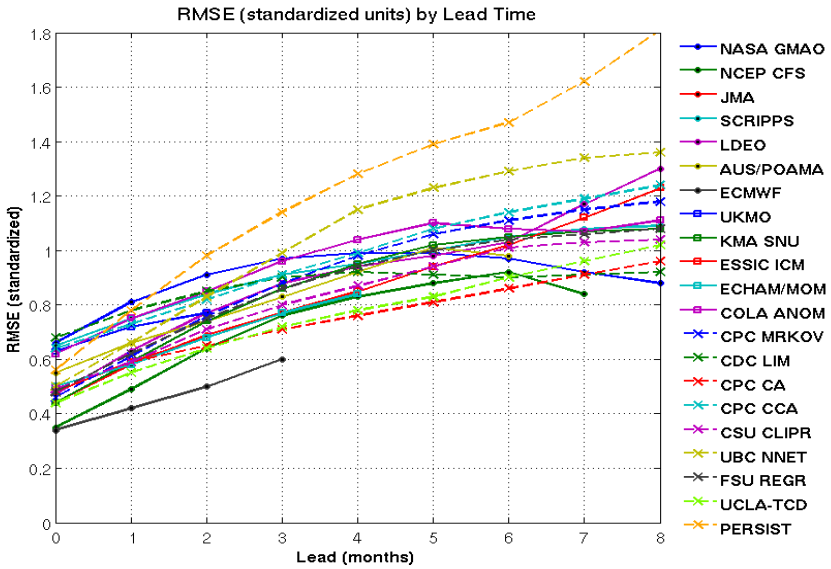

Figure 6 RMSE in standardized units, as a function of lead time for all seasons combined. Each line highlights one model. The 8 statistical models and the persistence model are shown with dashed lines and the cross symbol.

A likely reason that dynamical models are better able to predict ENSO through the time of year when transitions (dissipation of old events and/or development of new events) typically occur is their more effective detection, through the initial conditions, of new evolution in the ocean-atmosphere system on a relatively short (intramonth) time-scale—evolution that may go unnoticed by statistical models that use monthly or seasonal means for their predictor variables. Statistical models might be able to compete better against dynamical models if they used finer temporal resolution, such as weekly means.

Statistical models need long histories of predictor data to develop their predictor-predictand relationships. This need presents a problem in using the 3-dimensional observations in the tropical Pacific, such as the data from the Tao-Triton array, dating from the 1990s. (However, some subsurface tropical Pacific data do date back 10 or more years earlier in the eastern portion of the basin, and are available in the GODAS product.) This shorter data history precludes robust empirical definition of their predictive structures, and thus they are often omitted in statistical models. Although comprehensive dynamical models require a data history sufficient for verification and as a basis for defining anomalies, such a history is not basic to their functioning, and real-time predictions are able to take advantage of improved observing systems as they become available, potentially resulting in better initial conditions. While use of such crucial data suggests that dynamical models should be able to handily outperform statistical models, dynamical models have been burdened by problems such as initialization errors related to problems in data analysis/assimilation, and biases or drifts stemming from imperfect numerical representation of critical air-sea physics and parameterization of small-scale processes. As these weaknesses have improved, some comprehensive dynamical models have begun demonstrating their higher theoretical potential. This improvement will likely continue (Chen and Cane 2008).

b. Real-time Predictive Skill versus Longer-Period Hindcast Skill

Because 9 years is too short a period from which to determine predictive skill levels with precision, one reasonably might ask to what extent the performance levels sampled here could be expected to hold for future predictions. To achieve more robust skill estimates, a commonly used strategy is to increase the sample of predictions by generating retrospective hindcasts—"predictions" for past decades using the same model and procedures as in real time, to the extent possible. Cross-validation schemes are often used with statistical models, where varying sets of one or more years are withheld from the full data set, and the remaining years are used to define the prediction model which then is used to forecast the withheld year(s). In practice, there is

no comparable procedure applied in dynamical model development, and model parameter choices are often made using the same data used to evaluate skill.

Fourteen of the 20 models whose 9-year real-time forecast performance was discussed above (6 dynamical, 8 statistical) have produced hindcasts for the approximately 30-year period of 1981 (or 1982) to 2011. To assess the consistency of their skills during the longer period and the 9-year period of real-time predictions, the temporal correlation between hindcasts and observations is examined as a function of target seasons and lead time. Figure 7 shows a comparison of the correlation skills for the 9-year real-time predictions (as in Fig. 3) and the 30-year hindcasts for the subset of models having both data sets. Although the correlation plots are roughly similar, inspection shows generally higher hindcast skill levels for all of the models. Why do the hindcasts have higher skills? One explanation is that the 2002-2011 period may have been more difficult to predict than most of the longer period. Another explanation is that skills tend to be higher in hindcasts than in real-time predictions because the cross-validation designs may still allow inclusion of some artificial skill.

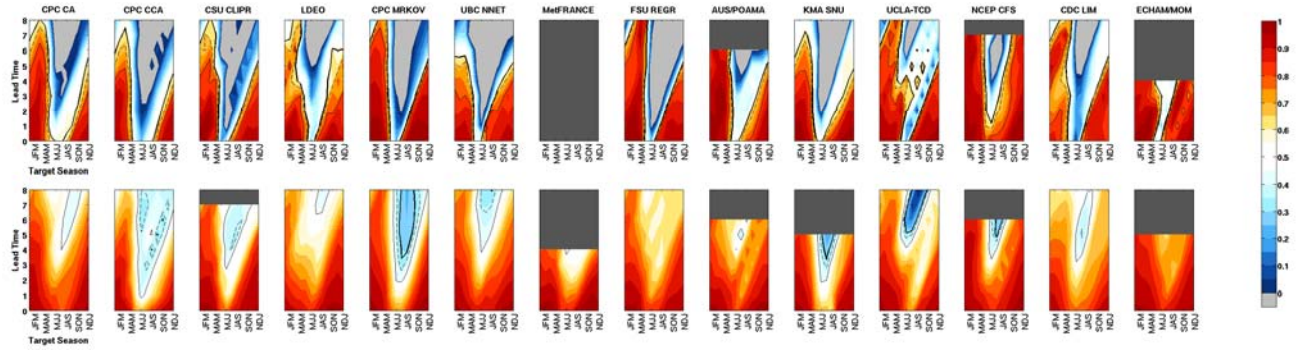

Figure 7 Temporal correlation between model forecasts and observations as a function of target season and lead time for (top) real-time forecasts (as in Fig. 3) and (bottom) hindcasts for the 1981-2010 period for models having long-term hindcasts. Thick solid contour shows the 90% significance level, dashed contour the 95% level, and thin solid contour the 99% level for sample sizes of 9 (top) and 30 (bottom).

To assess the relative difficulty of the recent 9-year period, the time series of uncentered correlation skills[2] of sliding 9-year periods, each phased 1 year apart, are examined for the 22 running periods within 1981-2010. The resulting time series of correlation are shown in Fig. 8 (top), for lead times of 3 months and 6 months. It is clear that for all models, and for both lead times, the 2002-2010 period, as well as the early to middle 1990s, posed a greater predictive challenge than most of the last three decades. As noted earlier in Fig. 1, one distinguishing feature of the 2002-2011 period is a lower amplitude of variability (no very strong events). The feature may be expected to reduce the upper limit of correlation skill by reducing the signal part of the signal-to-noise ratio. If the noise component remains approximately constant, and signal strength is somewhat restricted as during 2002-2010, then the correlation is reduced. The bottom inset of Fig. 8 (top) shows the 9-year running standard deviation of the observed Niño3.4 SST anomalies, with respect to the 1981-2010 mean. The correlation between the running standard deviation and the model average skill is about 0.8 for both 3- and 6-month lead predictions, confirming a strong relationship between signal strength and correlation skill.

The average of the anomaly of the 2002-2011 correlation with respect to that over 1981-2010 is approximately -0.14 (0.61 vs. 0.75) for 3 month lead forecasts and -0.23 (0.42 vs. 0.65) for 6-month lead forecasts. The -0.23 “difficulty anomaly” for 6 month lead forecasts is of greater magnitude than the deficit in skill of the real-time predictions during 2002-2011 compared with the approximately 0.6 skill level found in earlier studies, suggesting that today’s models would slightly (0.65 versus 0.60) outperform those of the 1990s if the decadal fluctuations of the nature of ENSO variability were taken into account.

[2] For the uncentered correlation, the 9-year means are not removed, so that standardized anomalies with respect to the 30-year means, rather than the 9-year means, are used in the cross-products and the standard deviation terms.

To examine the signal versus skill relationship with more precision, a 3-year time window is used in Fig. 8 (bottom), the bottom inset again indicating the running standard deviation. Within the 2002-2010 period, the subperiod of 2003-2007 is a focal point of low skill and low variability. The correlation between the 3-year running standard deviation and model average skill is again about 0.8 for both 3- and 6-month lead predictions, confirming a strong linkage between signal strength and correlation.

A second cause of the recent real-time predictions having lower correlation skill than the 30-year hindcasts is that using a period for which the verifying observations exist may permit inclusion of some artificial skill not available in real-time predictions. Attempts to design the predictions in a manner simulating the real-time condition (*e.g.* cross-validation) reduce artificial skill, but subtle aspects involving predictor selection often prevent its total elimination. Another impediment to the skill of real-time predictions includes such unavoidable inconveniences as delays in availability of predictor or initialization data, computer failure or other unforeseen emergencies, or human error. While this factor may seem minor, experience with the ENSO prediction plume has shown that such events occur more than once in a while.

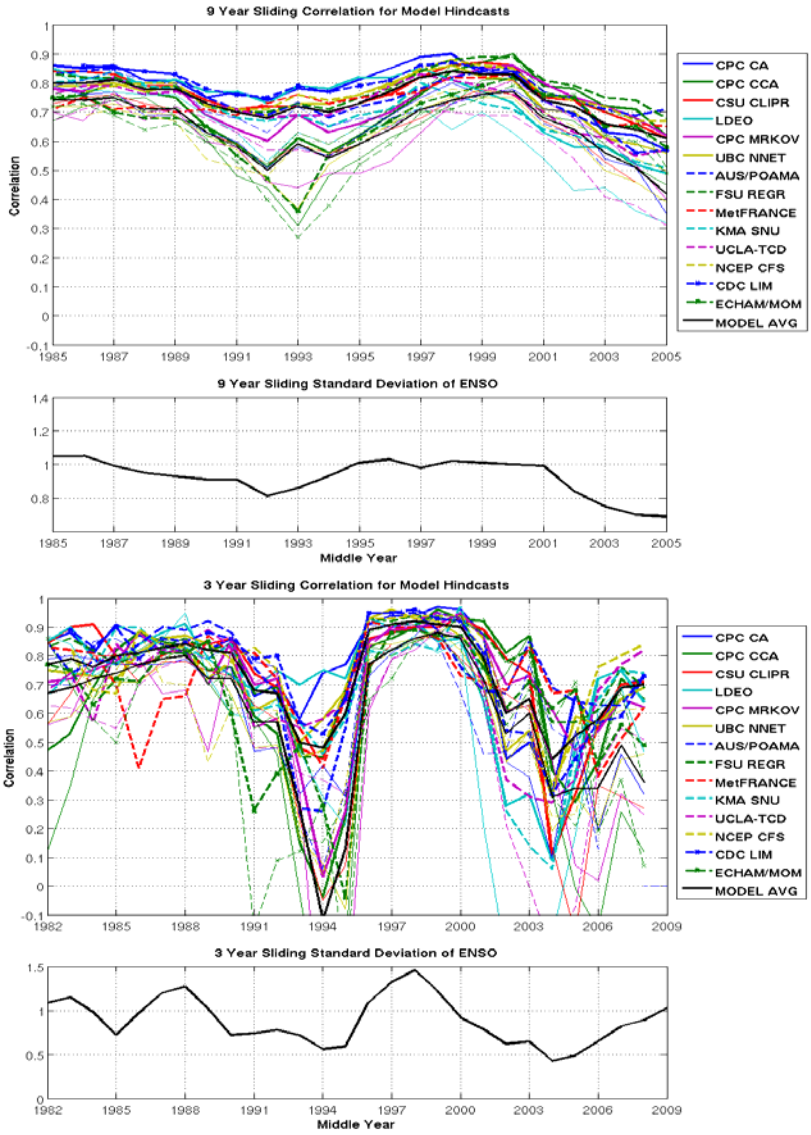

Figure 8 (top) Time series of uncentered correlations between predictions of given individual models and observations for sliding 9-year periods, phased 1 year apart, for the 21 or 22 running 9-year periods within 1981-2010. Correlations for forecasts at lead times of 3 months (thick lines) and 6 months (thin lines) are shown. Inset below main panel shows the standard deviation of observed SST anomalies for each sliding 9-year period, with respect to the 1981-2010 mean. (bottom) As in top, except for sliding 3-year periods for the 27 or 28 running 3-year periods within 1981-2010.

3. Concluding remarks

Verification of the real-time ENSO prediction skills of 20 models (12 dynamical, 8 statistical) during 2002-2011 indicates skills somewhat lower than those found for the less advanced models of the 1980s and 1990s. However, this apparent retrogression in skill is explained by the fact that the 2002-2011 study period was demonstrably more challenging for ENSO prediction than most of the 1981-2010 period, due to a somewhat lower variability. Thirty-year hindcasts for the 1981-2010 period yielded average correlation skills of 0.65 at 6 month lead time (slightly higher than the 0.6 found in studies from the 1990s), but the real-time predictions for 2002-2011 produced only 0.42. The fact that the recent predictions were made in real-time, in contrast to the partially hindcast design in the earlier studies, introduces another difference with consequences difficult to quantify but more likely to decrease than increase the recent

performance measures. Thus, based solely on the variability of 9-year correlation skills of the hindcasts within the 30-year period, ENSO prediction skill is slightly higher using today's models than those of the 1990s (0.65 versus about 0.6 correlation). Decadal variability of ENSO predictability can dominate the gradual skill improvements related to real advances in ENSO prediction science and models.

Unlike earlier results, the sample mean of skill of the dynamical models exceeds that of statistical models for start times between March and May when prediction has proven most challenging. The skill comparison by model type passes a field significance test for all seasons and leads collectively, at the p=0.03 level.

A likely reason for the better performance of dynamical than statistical models is a more effective detection and usage in dynamical models, through their initial conditions, of new information in the ocean-atmosphere system on a short (intramonth) time-scale—information that may not play a role in statistical models that use longer time means for their predictor variables. Statistical models may have potential for higher skill if their predictors were designed with finer temporal resolution. Statistical models need long histories of predictor data to develop their predictor-predictand relationships, but the valuable 3-dimensional observations across most of the tropical Pacific (e.g. from the Tao-Triton array) began only in the 1990s, precluding a robust empirical definition of their predictive relationships, and thus they are often omitted in statistical models, putting them at a relative disadvantage.

References

Barnston, A. G., and Coauthors, 1994: Long-lead seasonal forecasts – Where do we stand. *Bull. Amer. Meteor. Soc.*, **75**, 2097-2114.

Chen, D., and M. A. Cane, 2008: El Niño prediction and predictability. *J. Comput. Phys.*, **227**, 3525-3640.

Jin, E. K., and Coauthors, 2008: Current status of ENSO prediction skill in coupled ocean-atmosphere models. *Clim. Dyn.*, **31**, 647-664.

Livezey, R. E., and W. Y. Chen, 1983: Field significance and its determination by Monte Carlo technicques. *Mon. Wea. Rev.*, **111**, 46-59.

Ramseyer, G. C, 1979: Testing the difference between dependent correlations using the Fisher *Z*. *The J. of Experimental Education*, **47,** 307-310.

Wilcoxon, F., 1945: Individual comparisons by ranking methods. *Biometrics Bulletin*, **1**, 80–83. doi:10.2307/3001968.