Improving CFS seasonal prediction

WCRP Open Science Conference, October 24-28, 2011, Denver, CO

Anna Borovikov, Arun Kumar, Wanqiu Wang

Climate Prediction Center at NCEP/NOAA, Camp Springs, MD *email: Anna.Borovikov@noaa.gov*

Introduction

A seasonal forecast can be improved by various means: models can be better (for example, an increase in computer power can lead to finer resolution and more precise modeling of physical processes); initial conditions can be more accurate (either due to increase in data coverage and/or accuracy or improved assimilation techniques, for example); again, given resources, an ensemble size can be increased to give a better forecast and an error estimate.

The study presented here describes an attempt to improve the seasonal forecast by redesigning the existing ensemble so that it includes some selected members that were initialized earlier in time. This approach aims to gain in forecast skill at little extra computational expense.

A similar approach has been recently used by Meehl et all (2010) who studied the decadal variability of the Pacific using 30 member ensemble of 60-year long coupled model runs with perturbed initial conditions. The observations were simulated in a perfect model scenario. The criteria for member selection was quite straight forward - to "calculate the Euclidean distance from each ensemble member from the reference case in the ten year period from 1991-2000; the ensemble members with the lowest summed distance are chosen." The ensemble made up of the 9 best preselected members does follow the observations better than the full 30 member ensemble for a period of time at least comparable to the length of the training period (approximately 10 years). To test this idea's applicability to seasonal predictions we analyzed the monthly mean sea surface temperature (SST) data from 26 years of 9-month CFS forecasts with 15 ensemble members (Kistler et al., 2001).

Data and experiments

CFS forecasts from 1981 through 2006 were use for this study. Every month an ensemble of 15 members was integrated for 9 months. An example of the ensemble forecast plume is shown in the figure 1. Here the training period is chosen to be 3 months and the evaluation period is 6 months (the overlapping time for current and lagged ensemble members). Monthly mean SST anomaly is spatially averaged over the Niño3.4 region.

The current ensemble is shown in green, lagged ensemble members are blue and red, with the best 5 members of the lagged ensemble being red and the rest - blue. The corresponding ensemble means are shown with thick starred lines. Observations for training period are filled black circles, for evaluation period observations are white circles. The figure 2 shows only ensemble means for various forecast schemes for clarity. Root-mean-square error is used as a criteria to select the "best" lagged forecast ensemble members, always choosing five members with the minimum sum of the RMS error over the training months. The same criteria (RMS difference between the forecast and observations) was used to measure the forecast skill - sum over the target months during the evaluation period. The figure 2 corresponds to the setup when the training period is 3 month and therefore the evaluation period is 6 months. We have also studied the case of 2 months training and 7 months evaluation.

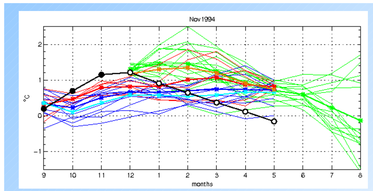


Figure 1: An example of the ensemble plume of the SST forecast, averaged over the Nino3.4 region. .

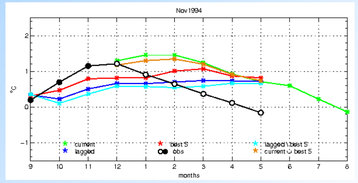


Figure 2: The same as figure 1, but only the mean curves are shown.

Evaluation

Focusing on comparing the current forecast and the current ensemble augmented with the 5 best lagged members we consider the following questions:

- Can we quantify the improvement?
- How can we assess the statistical significance of the difference in performance?
- Is the improvement due to a larger ensemble or a smart selection of lagged members? What if we just add as many members as possible?
- Is it worth including in the operational forecast?

Case studies

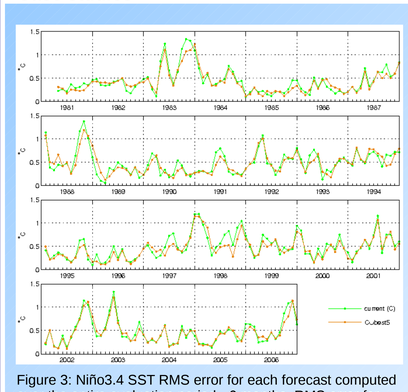


Figure 3: Niño3.4 SST RMS error for each forecast computed over the entire evaluation period - 6 months. RMS error for the current forecast is shown in green and the current + best5 lagged members is orange.

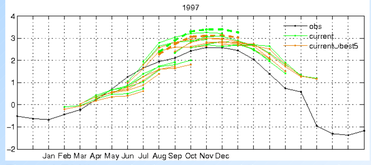


Figure 4:RMS error for the current and the augmented forecasts for every months of 1997. The forecast initialized in July is highlighted by thick dashed line.

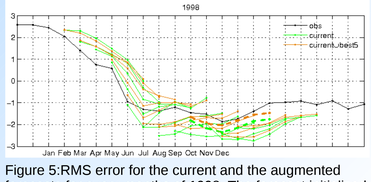
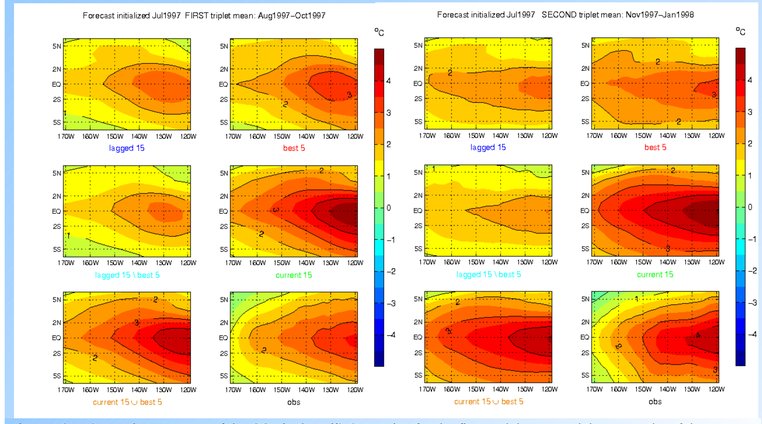


Figure 5:RMS error for the current and the augmented forecasts for every months of 1998. The forecast initialized in September is highlighted by thick dashed line.

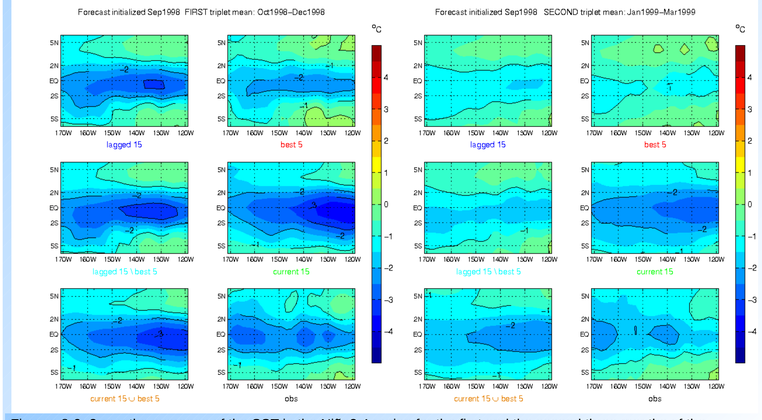
The figure 3 shows the SST RMS error averaged over the Niño3.4 region for the two forecast schemes computed over the entire evaluation period of 6 months. The error for the current forecast is shown in green and for the augmented forecast is in orange. Most of the time, the two forecast schemes preform similarly, however there are occasions when the difference appears to be quite noticeable. The figures 4 and 5 show the error for the every target month of the evaluation period for forecast starts in 1997 and 1998 years respectively. The two extreme cases, initialized in July 1997 and in September 1998, are highlighted by thick dashed lines.

1997 was a strong El Niño year, so there is a warming trend in the observations. Once the warming started, the current (green) forecast overestimated the rate of temperature increase, while the lagged ensemble members were still on the cooler side. Together they made the combined forecast (orange) appear closer to the observations. In 1998 the cooling takes place, and again, the forecast initialized at the beginning of the evaluation period, overestimated the trend. Inclusion of the lagged members helped to keep the combined forecast closer to the observations.

The figures 6-9 show the SST 2D fields for the two cases of July 1997 and September 1998. Plotted here are 3-months averages of the SST in the Niño3.4 region. The first three months of the evaluation period and then the second three months.



Figures 6-7: 3-months averages of the SST in the Niño3.4 region for the first and the second three months of the evaluation period started in July 1997.



Figures 8-9: 3-months averages of the SST in the Niño3.4 region for the first and the second three months of the evaluation period started in September 1998.

During the first three months of the July 1997 forecast (figure 6) the current forecast was too warm, the lagged one is too cold; the best 5 members were closer to the observations than the rest of the ensemble, yet the shape of the contours was not quite the same as in the observations. The current forecast, while overestimating the amplitude of the anomaly, better replicated its shape. The combined forecast was closer to the observations than the current. Further on (figure 7) the lagged ensemble, even the best 5 members, lost the resemblance to the observations. The current forecast was still too strong. A combination had the right amplitude at the maximum but too weak zonal gradient with respect to the observations.

The early half of the forecast initialized in September 1998 (figure 8) demonstrated the opposite extreme: the current forecast was too cold with respect to observations. The structure and the amplitude of the lagged ensemble was better. The addition of the best 5 lagged members to the current ensemble tamed the cooling especially during the first three months of the evaluation period. During the second half of the evaluation (figure 9), the best 5 lagged members still had a structure resembling the observations, although the amplitude of the anomaly was smaller. But the current forecast was still too cold, thus combining the current ensemble and the selected lagged members was beneficial to the forecast.

Probabilistic analysis

To evaluate how good is the choice of best lagged members in all the forecasts we check how the members that were the best during the training period perform during the evaluation period compared to the other possible choices of members from the lagged ensemble. We can analyze the distribution of ensembles created by adding not the best 5 members, but any 5 members, i.e. selected at random. In fact, one can learn the distribution of the random forecasts exactly by exhausting all the possible combinations of 5 members out of 15 from the lagged ensemble.

For each forecast initialization time we compute the mean error for all the ensembles constructed in such a way and the confidence intervals (50% and 95%). To get a sense how "the best" forecast is doing against the random case, we count how many times "the best" one is better than the mean of the distribution. Better in our case means smaller error.

We can also count the times "the best" forecast falls outside the confidence intervals. We can do this for different lead times to see how the skill of the forecast evolves over the evaluation period. Shown in the figure 10 is the deviation of the best forecast from the mean of the random forecasts and the 50% (shaded) and 95% (solid line) confidence intervals. The values have been calculated for the entire evaluation period of 6 months. The table contains the counts described above to illustrate the position of the best forecast (in the RMS error sense) with respect to the PDF of random forecasts.

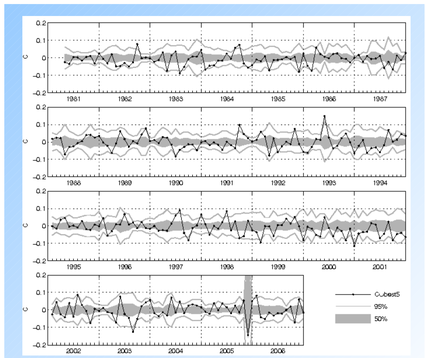


Figure 10: The deviation of the best forecast from the mean of the random forecasts and the 50% (shaded) and 95% (light solid line) confidence intervals.

lead (months)	1		2		3		4		5		6	
	<	>	<	>	<	>	<	>	<	>	<	>
mean	**200**	109	**193**	116	**191**	118	**189**	120	**178**	131	**181**	128
50% conf. int.	**159**	60	**135**	64	**130**	74	**120**	75	**108**	76	**113**	76
95% conf. int.	**55**	22	**41**	19	**37**	19	**28**	16	**31**	21	**25**	20

In bold are the values that are below the corresponding threshold, and the fact that they are always greater than their counterparts that are above the mean or upper bound of a confidence interval signifies that the choice of the best ensemble indeed makes it better in the probabilistic sense throughout the evaluation period.

Overall performance and statistical significance

The figure 11 shows the mean RMS error averaged over all forecasts (309 instances) vs the lead time. The 95% confidence intervals calculated by bootstrapping (Wilks, 2006, chapter 5) for each of the various forecast schemes for every target month of the overlapping evaluation period. Green color is of the regular forecast. It provides a reference for other ensembles. Blue squares show the error for the mean of the ensemble created by adding all 15 members of the forecast ensemble started 2 months earlier to a total of 30 members. Blue circles - the error for the mean of the ensemble made up of all the current members and all the members of an ensemble started 3 months earlier to a total again of 30 members. Blue diamonds show the error for the mean of a 45-member ensemble which includes all the current members, all the members trained for 2 months and all the members trained for 3 months.

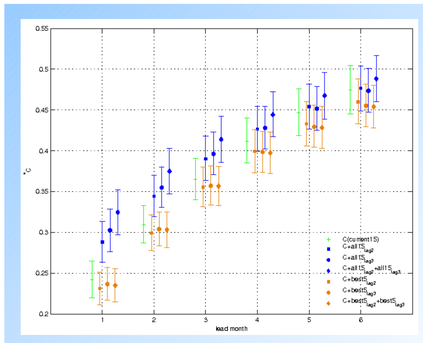


Figure 11:The mean root-square error averaged over all forecasts (309 instances) vs the lead time. The 95% confidence intervals calculated by bootstrapping for various ensembles for the overlapping evaluation period.

At the beginning of the evaluation period, the larger ensembles compare poorly to the current forecast, yet towards the end, the blue dots do as well or slightly better that the green ones. This should not be surprising - early on more skill is due to a good initialization of the ensemble, while later on, the improvement may be attributed mostly to the increased ensemble size. The orange squares, circles and diamonds represent the combinations of the current ensemble with the best five members from the ensembles initialized 2 and 3 months earlier. There is no loss of skill in the beginning of the forecast, and towards the later dates, the errors of the combined ensembles remain smaller than that of the current forecast.

Conclusion

The ensemble augmentation technique presented in this study leads to a robust albeit small improvement in the SST forecast over 6 to 7 months range. An advantage of the proposed approach is its low computational cost. The training procedure can be done off-line and the selected lagged ensembles members have to run for additional 2 to 3 months (in the case of 9 months forecast). Thus an improvement can be gained at little additional expense.

Simple non-discriminating training shows promise of the approach. Not mere ensemble size, but better ensemble. More precise training techniques can produce better results, for example, clustering the ensemble members based on their proximity to the observations.

References

R. Kistler, E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu, V. Kousky, H. van den Dool, R. Jenne, and M. Fiorino, 2001: The ncep-ncar 50-year reanalysis: Monthly means cd-rom and documentation. *Bull. Amer. Meteor. Soc.*, **82**, 247–268.

G. A. Meehl, A. Hu, and C. Tebaldi, 2010: Decadal predication in the pacific region. *Journal of Climate*, **23**, 2959–2973.

J. J. Ploshay and J. L. Anderson, 2002: Large sensitivity to initial conditions in seasonal predictions with a coupled ocean-atmosphere general circulation model. *Geophysical Research Letters*, **29(8)**.

D. S. Wilks, 2006: *Statistical Methods in the Atmospheric Sciences*. International Geophysics Series, Vol.91. Elsevier, second edition.