

## Performance Characteristics of Forecasts Based on a Lagged Ensemble

David A. Unger

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

### 1. Introduction

The Climate Forecast System Version 2 (CFSv2) is a coupled atmospheric-oceanic global climate model run by the National Centers for Environmental Prediction (NCEP). The CFSv2 is run operationally each 6 hours and makes forecasts out to around 6 months lead time. Uncertainty is estimated by a forecast ensemble composed of recently available initial times and aligned according to the target period in question. Thus, an ensemble forecast set is formed from a series of runs lagged backwards in time from the run based on the most recent data, and is hence referred to as a lagged ensemble (Hoffman and Kalnay 1983). This contrasts with the approach commonly used for shorter range forecasting of producing an ensemble forecast from multiple runs initialized from perturbed initial states based on the most recent data.

For lead times typical of climate forecasts, usually measured in months, the forecast benefits of a larger ensemble more than compensates for losses due to the inclusion of negligibly less skillful members from runs a few hours or days old. However, there is eventually a point where the trade-off between ensemble size and loss of skill due to the inclusion of older members becomes important. At some point the inclusion of less accurate ensemble members from older runs will begin to detract from the information available from members based on more recent data. An objective procedure to build a lagged ensemble set is proposed in this paper. This method is based on the skill characteristics of the candidate ensemble members. The result is an objective weighting method for ensemble members from variously lagged initial times.

### 2. Data

Sea Surface Temperature (SST) forecasts for the Nino 3.4 region from the CFSv2 will be used to test the method. Hindcast data initialized every 5<sup>th</sup> day from 1982 – 2010 is available from the CFSv2. The hindcast data was adjusted for a discontinuity in forecast performance beginning in late 1998. A simple bias correction, stratified by initial time and lead, was applied to the earlier forecasts (1982-1998) to make their mean bias similar to the recent forecasts (1999-2010). Forecasts are expressed as anomalies relative to the 1981-2010 Nino 3.4 climatology.

The 4 ensemble members obtained from initial times 6 hours apart on any given day were assumed to be equal in skill and always were grouped together, treating them as if they were perturbations from a daily initial state. A lagged ensemble set was built from the series of 4-member daily runs available from CFSv2 hindcast data going backward in time from the run closest to the end of each calendar month, defined as lag 0. Members from 5-days prior to lag 0 were labeled as lag 5, 10 days prior as lag 10 and so on. The forecasts were for the monthly mean Nino 3.4 SST for the month following the lag 0 forecast (defined as lead 1), the subsequent month (lead 2), and successive months out to lead 6. Up to 20 4-member ensemble sets were offered for potential inclusion in a lagged ensemble (out to lag 95), adjusting the lead time of earlier runs as necessary to align the forecast relative to lag 0.

### 3. Methods

#### *a. Ensemble Regression*

A regression procedure specifically designed for ensemble forecasting was used to process the CFSv2 forecasts. Ensemble regression (Unger, *et al.* 2009) estimates the expected values of the coefficients of a regression equation relating the closest member of a set of ensemble solutions to the observation, together with its expected error. In addition to the usual requirements for the appropriate application of linear

regression and its error estimate (a reasonably linear fit, Gaussian distributed residual errors, case to case independence of errors, etc.), ensemble regression requires a reasonable a priori estimate of the probability that each ensemble member will be closest to the observation, usually taken to be equal for all members. Once this estimate is supplied, a regression equation can be derived that gives theoretically expected values of the coefficients, together with an expected value of the standard deviation of the errors, assumed to be Gaussian, about the regression estimate. This equation, together with the error estimate, is applied to each ensemble member.

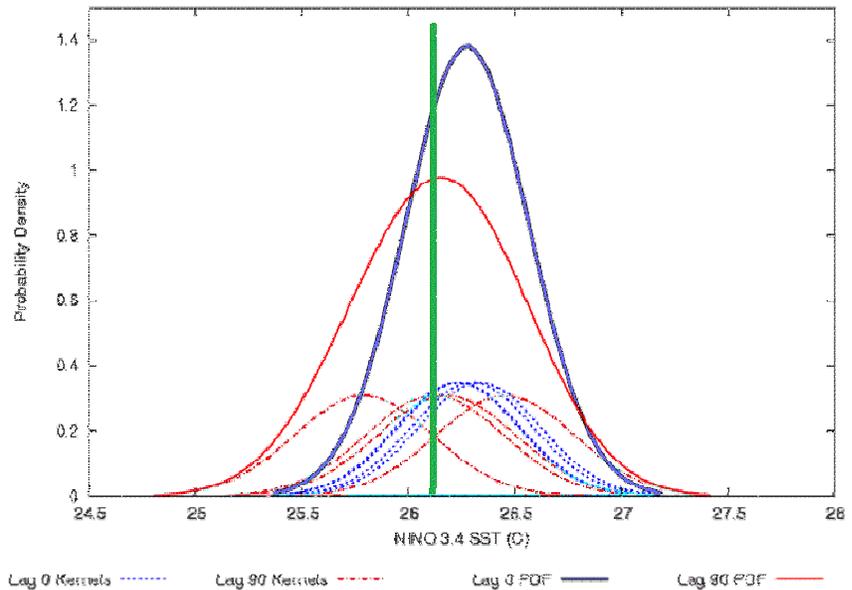
Ensemble regression was used to calibrate the forecasts for an ensemble consisting of the 4 runs available for a calendar day. Each of these ensemble members were assumed equally likely to be closest to the observation (best member). In its application, ensemble regression produces an estimate of the forecast probability density function (PDF) of the combined ensemble. It is similar to a PDF obtained from a Gaussian kernel density estimate (Roulston and Smith 2003), except that the kernels are centered around the regression estimates of each ensemble member, rather than their original values. The standard deviation of the kernel distributions are specified by the regression procedure.

Regression equations were derived for the predicted monthly mean Nino 3.4 SST from each of 73 initial times (every 5 days referred to here as pentads) throughout the year. The equations for each pentad were based on the 29 cases for that initial time in the 1982-2010 period. A separate equation was derived for each monthly lead time. Results reported here are based on dependent data (no cross-validation).

Figure 1 illustrates the forecast PDF from two separate groups of ensemble forecasts. This is a forecast for the monthly mean Nino 3.4 SST for July, 2010. One forecast (indicated by blue) was initialized on June 30, 2010, only 1 day prior to the start of the valid period. The solid line is the combined PDF from the 4 kernel distributions representing individual members, shown by dashed lines. The forecasts at this lead were highly skillful, indicated by closely grouped, narrow kernels. A forecast made 3-months earlier (April 1, lag 90) for the same target period is shown in red. Note that the expected errors around the best member (illustrated by kernels) are only slightly less skillful than the shorter lead time as evident from the kernel width. This indicates that most of the difference in forecast uncertainty is accounted for by the increased amount of spread compared to the earlier run. The green vertical line shows the observed mean SST for July 2010.

*b. A procedure for weighting a lagged ensemble*

Figure 1 also illustrates a principle for optimizing a lagged ensemble. The PDF from the ensemble regression in the vicinity of the observation provides a measure of how closely the forecast and observation match. The PDF from each candidate ensemble forecasts (one for each lag) is used to provide a measure of which was “best” in a given case. The ensemble set with the highest PDF value at the observation is, by inference, the one that is most likely to be correct. The regression calibrates



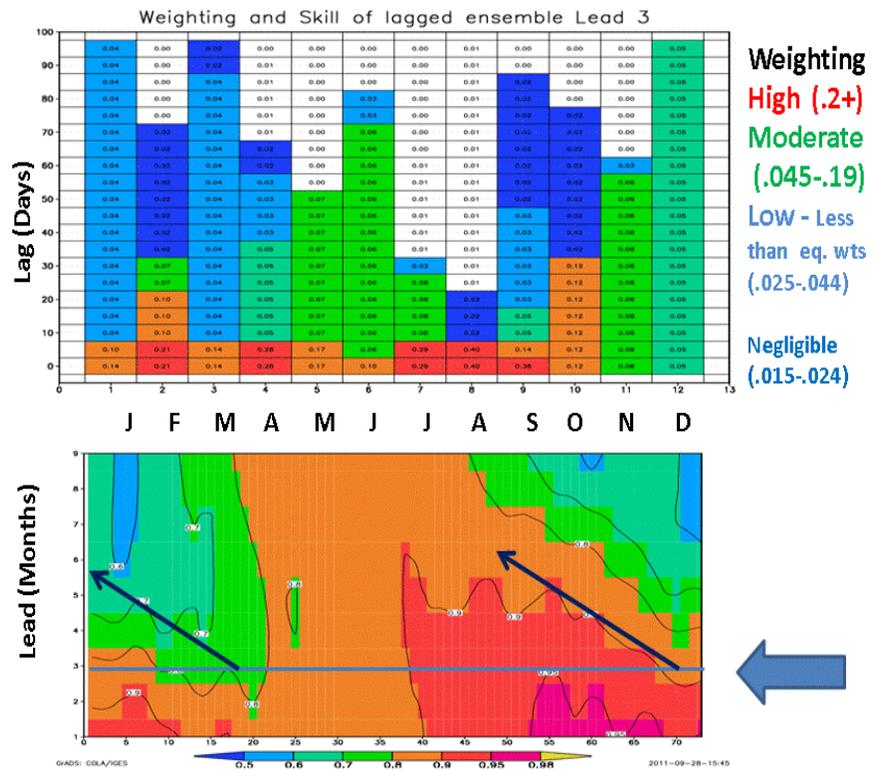
**Fig. 1** Forecast PDFs based on ensemble regression from CFSv2 forecasts of Nino 3.4 SST from two initial times. Blue lines represent forecasts of monthly mean Nino 3.4 SST initialized on June 30, 2010 valid for July, 2010. Red lines are for the Lag 90 forecast (initialized April 1). Kernels represent individual members and are shown by dashed lines, with the combined PDF in solid. The green vertical line represents the observed SST.

the entire ensemble in relation to its performance on hindcast data. So, in this case, the single best member of the 8-member ensemble formed from the lag 0 and lag 90 runs happened to be from the lag 90 set. However, the regression processing accounts for both the spread and the skill to better match the forecasts and observations. The regression calibrated PDF indicated that, considering its spread, and the skill of CFSv2 forecasts from similar leads on the hindcast data, the more recent run, in fact, was the better match to the observation.

The potential ensemble system consisting of the 20 most recently available 4-member lagged ensemble forecast sets were evaluated on the dependent (hindcast) data by passing through the data twice, once to derive the ensemble regression equations, and once to evaluate forecasts and produce a separate PDF for each lag. The ensemble set that produced the “best member” was assumed to be the one with the highest PDF in each case. The probability that each of the 20 candidate ensemble sets produced the best member (P(best)) was estimated from the fraction of times out of the 29 cases (one per year) that a given lag was best. The results were smoothed under the assumption that P(best) decreases with increasing lag time. If this was not so, the older members were assumed equally likely to produce a best member than any more recent ensemble set with a lower P(best). The P(best) for each lag was then re-evaluated with the combined ensemble, effectively averaging the P(best) values over any inconsistent lead times. This way P(best) of an older run was assured to be equal to or lower than a more recent run.

**4. Results**

The top panel of Fig. 2 shows P(best) for lagged ensemble forecasts for SSTs initialized near the end of each of 12 calendar months, for a lead-3 forecast. The vertically stratified bins (columns) correspond to the initial month, and the boxes along the vertical (rows) correspond to the lag relative to the most recent run. The bottom row represents the four member ensemble closest to the end of the month in the corresponding column. Each higher row represents an ensemble set initialized 5 days earlier than the row below. The inset numbers show P(best) for that ensemble set, objectively determined from the hindcast data as described above. P(best) is an appropriate weighting for that ensemble set in ensemble regression theory. The final forecast of a weighted ensemble would be formed from a kernel density approach similar to that shown in Fig. 1 except with the area of each of the 4 kernel proportional to 1/4 of the total weight assigned to that lag.



**Fig. 2** Weighting for each lag (P(best)) stratified by initial month and lead time (top panel), with Hovmoller plot of regression skill by lead and initial time in pentads (bottom panel). Inset values in the bottom panel represent the correlation between the ensemble mean and the observation for a forecast for monthly mean Nino 3.4 SST from the CFSv2. See text for further explanation.

For example, a forecast for July, initialized at the end of June (column 7, bottom row), was the “best” of any of the 20 lagged ensemble sets offered 29 percent of the time. This run was initialized on June 30 of each

year of hindcast data. The run initialized 10 days before that, on June 20, was best 8 percent of the time. The oldest run to contribute significantly to lagged ensemble was from lag 30 (initialized on May 31) and accounted for only 3% of the best members (was best once out of the 29 year sample). Members initialized at lags between 35 and 55 days only sporadically contributed to the combined ensemble, accounting for trace weighting (1%), and no run 60-days or older contributed to the ensemble at all. Colors indicate the relative contribution, with red signifying a weighting significantly higher than an equally weighted 20-set ensemble. Green indicates moderate weighting, and blue indicating significantly less weight than would be expected for equal weights among candidate ensemble sets.

The bottom panel of Fig. 2 reveals the possible explanation for the ensemble weighting. It shows the Hovmoller plot of the skill of a regression equation based on the ensemble mean (an important component of the ensemble regression). The 73 initial pentads are along the horizontal axis, with integer lead time in months relative to the calendar month of initialization shown on the vertical. The correlation coefficient of the regression relationship is contoured within the diagram. Even with smoothing, some evidence of a saw-tooth pattern is visible in the plot, indicating that the skill is higher for runs initialized late in the month than those initialized earlier on. The horizontal line illustrates the lead time (3-month) used for the ensemble weighting diagram on the top panel. The arrows indicate the direction in which the lagged ensemble is built for runs initialized in March and December. The skill of most recent run is shown near the tail of the arrow, with older runs represented toward the arrow's head.

When the skill of the ensemble set falls off rapidly with increasing lead, as indicated by skill decreasing along the arrow towards its head, the objective ensemble weighting procedure heavily front loads the lagged ensemble set. The weight drops rapidly and becomes trivial even after a few lags. This is common around the time of the spring predictability barrier indicated by areas of strong gradients on the diagram. On the other hand, skill in predicting Nino 3.4 SST does not depend much on lead time for runs initialized late in the year. The skill at the head of the arrow (oldest runs) is not much different from the newest runs. The lagged ensemble at this time of year is fairly evenly weighted, in spite of lags of 30 days or more. Thus, late in the year a larger ensemble group is adding information to the system, compared to a smaller ensemble consisting only of more recent runs. This is not the case in the boreal spring, where the bulk of the weight is on the most recent members, and older members are trivially weighted at best. At this time of year, the newer runs are more skillful than older ones, and the ensemble members from older runs provide little if any information not represented by the more recent runs.

## 5. Conclusion

The procedure described in this paper provides an objective way to analyze the information in a time-lagged ensemble. It can be used to help determine when ensemble members from older runs can meaningfully contribute to an ensemble consisting of members from more recent runs. The method provides an objective weighting procedure for the lagged ensemble forecasts, and provides results consistent with expectations based on forecast skill. This method is not restricted to lagged ensembles, since it can be used as a basis for combining multi-model ensembles as well. Future work will focus on evaluating the skill of this procedure in a fully cross-validated framework.

## References

- Hoffman, R., N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus*, **35A**, 100-118.
- Roulston, M. S., and L. A. Smith, 2003: Combining dynamic and statistical ensembles. *Tellus*, **55A**, 16-30.
- Unger, D. A., H van den Dool, E. O'lenic, and D. Collins, 2009: Ensemble Regression. *Mon Wea. Rev.* **137**, 2365 – 2379.