HMOS Verification Exercise
2nd RFC Verification Workshop
Salt Lake City, UT
11/19/2008

Goal: compare the IVP and EVS capabilities and interpret the forecast verification results for an HMOS case study using the single-valued flow forecasts (used as input in HMOS), the HMOS ensemble mean flow forecasts, and the HMOS ensemble flow forecasts.

1. Display data

IVP Time Series Plot (slide #4)

Notice the periods with missing values.
Get an estimate of the number of events above flood level.

It is easier to view the observed time series in IVP itself, not from this static image (the time period is very large and the number of forecast times series as well) since you can zoom in. Here we need to assume that when several forecast values are above the flood level, a flood actually occurred. It is also difficult to see if the flood event lasted for several time steps or if several flood events occurred in a few days. However there are roughly 8 flood events. It means that the sample size is very small to compute verification statistics only for flood events. It makes more sense to define other threshold values to analyze large events, e.g. the 90th percentile from the observations record.

What variable is plotted on the Y axis?

Instantaneous discharge in cfs (not kcfs as stated on the graphic), for all forecast time series as well as the observed time series.

EVS Box Plots with time (slides #5-12)

What variable is plotted on the Y axis?

Discharge forecast error (forecast - observed) in cfs, which is computed for each ensemble member and for a given lead time. Then the distribution of these ensemble forecast errors is plotted with box and whiskers (Minimum, 10th percentile, 20th percentile, …80th percentile, 90th percentile, Maximum).

Comparing the boxes on slides #5-8, what can you say about the spread of the HMOS ensembles?

The spread of the ensembles increases with lead time since the uncertainty is larger. Generally the boxes include the zero error line, indicating that the ensemble forecasts captured well the observed values. There are a few missed forecasts, for which the whole box is below the zero error line, especially for longer lead times.

From slides #9-12, what can you say about the error in the HMOS ensemble means?

The error in the ensemble means increases with lead times, with more points further away from the zero error line and larger maximum errors. For shorter lead times, there is no strong tendency for under- or over-forecasting since one of the goals of the HMOS system is to correct for bias in the single-valued forecasts and provide unbiased ensemble mean values.

<u>IVP Time Series Plot (slides #13-15)</u>

What can you say about the performance of the HMOS ensemble means vs. the single-valued forecasts for the 3 different lead times?

The HMOS ensemble means tend to perform better than the single-valued forecasts with more points closer to the diagonal line; there is less over-forecasting for the lead hours 6 and 24 (e.g. for lead hour 24, the red points with forecast > flood and obs. < flood, compared to the fewer blue points), and less under-forecasting for lead hour 48.

<u>EVS Box Plots with observed value (slides #16-23)</u>

What is the difference between the graphics on slide #16 vs. slide #5?

The variable on the X axis: it is the observed flow on slide #16 vs. the time (starting from the first day of the verification period) on slide #5. Plotting with the observed values on the X axis is very useful to detect any conditional bias in the forecasts (e.g., positive bias increasing with observed value).

By comparing slides #16-19, what can you say about the conditional bias of the ensembles?

There is no clear conditional bias for lead hour 6. For the 3 longer lead times, there is a tendency for increased under-forecasting with larger events (observed values above ~25 kcfs). This conditional bias is very large for lead hour 72, with boxes well below the zero error line.

From slides #20-23, what can you say about the conditional bias of the ensemble means?

There is a tendency to under-forecast the ensemble means for higher events for the 3 longer lead hours; this conditional bias increases with lead time.

2. Error verification metrics

<u>IVP Correlation Plot (slide #24)</u>

What can you say about the correlation coefficients for the 3 sets of forecasts (ensemble mean, single-valued and persistence)?

The HMOS ensemble means (EM) and the single-valued forecasts (SV) have very similar Correlation Coefficient (CC) until lead hour 84, and from then HMOS EM perform worse. Correlation drops faster after lead day 2. The 2 sets of forecasts beat persistence forecasts at all lead times after lead hour 6.

<u>EVS Correlation Plot (slide #25)</u>

The HMOS ensemble mean forecasts were verified for 2 subsets based on 2 different thresholds: 5,297 cfs, which corresponds to the 90th percentile of the observations, and 32,100 cfs, which corresponds to flood level.
What can you say about the correlation coefficients for these 2 forecast subsets?

The Correlation Coefficient (CC) decreases with lead time. The CC is quite high for the first 2 lead days and then drops quickly to reach the zero line. The larger variations in the blue curve for the higher threshold value are due to the small sample size (just a few flood events leading to erratic statistical results). This reflects the skill in the singe-valued forecasts. Rapid decrease in forecast skill could be attributed to the basin's size and memory.

<u>EVS Mean Continuous Rank Probability Score (MCRPS) Plots (slides #26-27)</u>

The MCRPS corresponds to the Mean Absolute Error (MAE) for single-valued forecasts. How do the MCRPS values compare against each other for the HMOS ensembles and the HMOS ensemble means and for the 2 forecast subsets?

MCRPS increases with lead time and is higher for the flood level threshold, as expected. MCRPS is lower for the HMOS ensembles compared to MAE at all lead times and for the 2 subsets; this shows the added value of the ensemble forecasts, compared to the HMOS ensemble means.
For example, the values at lead hour 120 are:

- *for the 90th percentile threshold, MCRPS ~ 12 kcfs, MAE ~ 15 kcfs;*
- *for the flood threshold, MCRPS ~ 39 kcfs, MAE ~ 43 kcfs.*

<u>EVS Mean Error (ME) Plot (Slide #28)</u>

What can you say about the ME values for the 2 subsets of forecasts?

The ME values are small for lead hours 6 and 12 and then negatively increase with lead time. This under-forecasting bias is very large for the flood event subset. Again the very small sample size for this flood event subset leads to erratic statistical results.

IVP Error Plots (Slide #29)

What can you say about the ME and RMSE values for the 3 sets of forecasts?

Until lead hour 66, the HMOS ensemble means (EM) and the single-valued forecasts (SV) have very similar RMSE, the EM forecasts being slightly better than the SV forecasts. After lead hour 66, the EM RMSE increases with lead time whereas the SV RMSE does not vary much. The 2 sets of forecasts beat persistence forecasts at all lead times.
For forecast bias, the EM forecasts have the smallest bias values at all lead times (which is one of the goals of the HMOS system); the positive bias increases significantly with lead time after lead hour 48. The SV forecasts have a small positive bias for lead day 1; then the negative bias increases significantly with lead time. The SV bias is significantly higher than the persistence bias after lead hour 36.
The increase in RMSE at longer lead times for HMOS EM forecasts compared to SV forecast is driven to some extent by producing symmetric errors for the HMOS forecasts to reduce bias in the SV forecasts.

What can you say when comparing the ME values with slide #28?

The ME values in this IVP graphic are relative to the whole set of forecasts, whereas the ME values in the previous EVS graphic are relative to forecast subsets. The differences in these 2 ME curves show the conditional bias in the forecasts: under-forecasting bias for larger events.

3. Conditional verification metrics

EVS ROC Plots (Slide #30-37)

For both HMOS ensemble means and ensemble forecasts, ROC was computed for 2 events: flow > 5297 cfs and flow > 32210 cfs.
What can you say about the variations of ROC with lead time for the 2 events?
How does ROC for HMOS ensembles compare to ROC for HMOS ensemble means?

For lead hour 6, the HMOS ensembles have the same ROC curve for the 2 events, which is very close to the perfect ROC curve (which is the curve joining the points (0,0), (0,1) and (1,1)).This means that HMOS forecasts can discriminate well between the two alternative outcomes for each of the 2 events, thus they show good resolution.
The forecast resolution decreases with lead times for both events since the ROC curves get closer to the diagonal line. The degradation is stronger for the flood event, as expected. At lead hour 72, the resolution of the flood event is very poor (i.e. the blue ROC

curve is close to the diagonal line); the resolution for the lower threshold is significantly better.

The ROC curves for the HMOS ensemble means are based on only one point, corresponding to the Probability of Detection and Probability of False Detection for the given event. For all lead times, the ROC curves show similar or better event discrimination with the HMOS ensembles vs. HMOS ensemble means; the gain in the forecast resolution increases with lead time. This shows the added value in the ensemble forecasts vs. the single-valued forecasts for resolution.
The computation of the ROC area (the area under the ROC curve) will facilitate the comparison between different ROC curves and is a common summary statistic for forecast resolution.

EVS Cumulative Talagrand Plots (Slide #38-41)

The cumulative diagrams were produced for 2 subsets of forecasts: observed flow > 5,297 cfs and observed flow > 32,210 cfs.
What can you say about the variations of the Talagrand curves with lead time for the 2 forecast subsets?

For lead hour 6, the HMOS ensembles are very reliable for the lower threshold since the red curve is very close to the diagonal line, with a slight under-spread issue of the ensemble distribution. For the flood threshold, the ensembles show a significant lack of spread since the blue curve is below the diagonal line. The last point for a probability window of 1 (for which the fraction of observations is 0.8) shows that 20% of the time the observations are completely above the forecast distribution. The first point for a probability window of 0 (for which the fraction of observations is ~0.13) shows that 13% of the time the observations are completely below the forecast distribution.

For the other lead times, the lack of spread in the ensemble distributions is increasing with lead time, being worse for the flood threshold. For the flood threshold, for lead hours 24 and 48, the observations fall above all the ensemble forecast values more than 50% of the time; for lead hour 72, it is the case 85% of the time. This reflects the difficulty at longer lead times of predicting potential flood events with ensemble forecasts when the deterministic flow forecasts have little skill in predicting flooding.