24WAF20NWP 7.3

The performance of high-resolution WRF ensemble QPF during heavy winter precipitation events in California

Edward I. Tollerud[1], Tara Jensen[2], Isidora Jankov[4], Huiling Yuan[3], John Halley Gotway[2], and Paul Oldenburg[2]

[1]*NOAA Earth System Research Laboratory (ESRL), Boulder, Colorado*
[2]*Developmental Testbed Center and Research Application Laboratory, National Center for Atmospheric Research, Boulder, Colorado*
[3]*School of Atmospheric Science, Nanjing University, Nanjing, Jiangsu, P.R. China*
[4]*Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, and Earth System Research Laboratory (ESRL), Boulder, Colorado*

1. Introduction

A series of annual winter forecast exercises over vulnerable watersheds of California and Nevada have been undertaken as part of the NOAA Hydrometeorology Testbed (HMT) to assess the potential for improved quantitative precipitation forecasts during heavy precipitation events. For the most recent exercise, a nine-member high-resolution (9 km) ensemble WRF forecast system produced forecasts during a week-long period of heavy and persistent rainfall in Northern California. An online verification system funded by the USWRP and implemented at the Developmental Testbed Center (DTC) concurrently provided quantitative assessment of the performance of the individual ensemble members, the simple ensemble mean, and (as baseline) a GFS deterministic run with nominal resolution of 40 km. The quantitative precipitation forecasts are evaluated in two ways. First, qualitative comparison of several different traditional verification scores are presented for selected episodes to illustrate the several models' individual characteristics in different meteorological scenarios. Second, the verification scores are aggregated over a several-week period to describe overall performance and to provide a reasonable basis for comparison of the WRF ensemble with GFS forecasts. For each set of results, attention is paid to the impact and implications of different spatial resolution to the verification results.

2. The 2009-2010 HMT Winter Exercise

Domains were selected for the winter exercise that included a large area covering most of California and Nevada and extending several hundred km westward into the Pacific Ocean (Fig. 1). Eight ensemble member forecasts were produced in the large domain using both ARW and NNM cores of the WRF model initiated with several randomly-selected GFS ensemble members for boundary conditions. Forecasts were output every three hours up to 5 day lead times. The spatial resolution of this domain was approximately 9 km. An ensemble mean was produced from these members, and a coarser-resolution GFS forecast was included in the verification for base-lining. In addition, forecasts within a smaller nested domain were produced, and another domain with high temporal resolution (1 hr) was added for shorter duration forecasts. Verification results presented here are for the full domain.

To monitor forecast performance during the exercise, a real-time website was established to provide up-to-date and retroactive verification statistics for the 9 WRF ensemble members. This system allowed multiple scoring options including standard scores (equitable threat; false alarm; RMSE; bias; etc.) for runs at constant

initialization time and constant valid times, as well as object based techniques that keyed on quantitative precipitation forecasts. In addition, summary score statistics were routinely displayed for the previous 30 day period to gain a sense of past model performance. One of the innovative features of the system was the opportunity to select from a choice of verification datasets (e.g., Stage IV grids at 6h accumulation periods, and Stage IV and gages at 24h periods) and regions (individual watersheds and the California Nevada River Forecast Center domain). In this paper we present a first comparative assessment of WRF ensemble model performance and show results that reveal some impacts presented by the choice of data. Since baseline GFS model simulations (at approximately 40 km resolution) were also verified, it is possible to compare verification results that proceed purely from resolution differences.

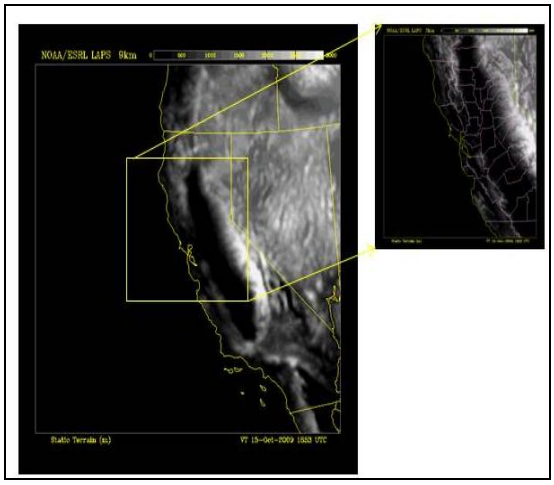

Fig. 1. The full and inner nested WRF ensemble domains during the HMT West winter exercise.

Many of the results shown here are from stormy periods in January 2010. During the week of 17-21 January, in particular, several storms moved onto the northern and central California coast resulting in heavy precipitation in most of the coastal mountains and the Sierra Nevada Mountains. Precipitation observations at operational gauge sites on January 20-21 (Fig. 2) illustrate the pattern that persisted during this episode.

The cumulative rainfall totals during 5 days of this episode (Figs. 3 and 4) at the observation site denoted by the large circle in Fig. 2 illustrate both the magnitude of the series of storms and the diagnostic challenges presented by ensemble forecasting systems. No clear pattern emerges in Fig. 3 for the relative accuracy or tendency exhibited by any of the individual members, but the ensemble mean has apparent value compared to the cumulative GFS forecast (Fig. 4).

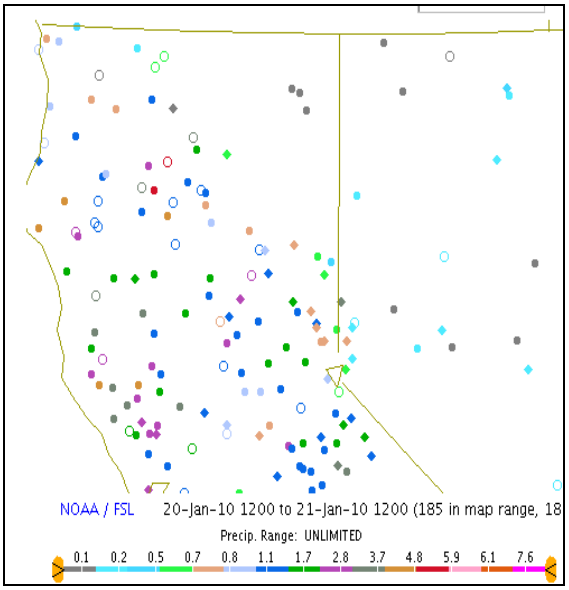

Fig. 2. Rainfall in inches at rain gage sites for the 24h ending 1200 UTC 21 January 2010.

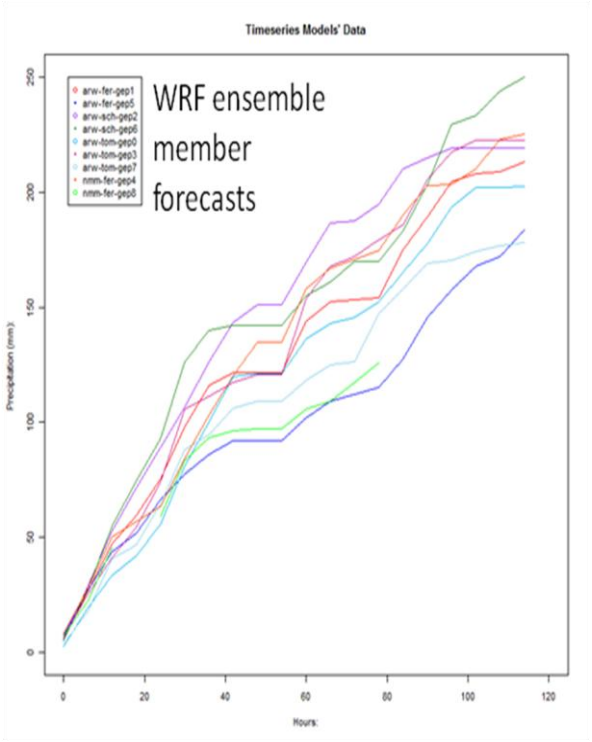

Fig. 3. Cumulative rainfall at six hour intervals at the WRF domain grid center point nearest the Alta observation site in the Sierra foothills. Forecasts from 9 WRF ensemble members are shown. Time series commences at 1200 UTC 17 January 2010.

3. Summary Statistics

The above general assessment of GFS forecasts as compared to the WRF ensemble mean was limited to one set of rainfall events. In Fig. 5, the results are generalized to a thirty day bulk analysis during the entire month of January. Conclusions to be drawn are consistent with the previous analysis, in that the ensemble mean GSS is generally larger (better) than that of GFS forecasts. This comparison is particularly evident for larger rainfall thresholds, an indication that the higher resolution of the ensemble system is especially important for heavy rainfall events. The size of the bars is a sign that differences are not significant, especially that between the ensemble members.

Fig. 6 presents a similar comparison, with the difference that the groupings are for lead time categories. It is clear that the same pattern emerges: the higher-resolution WRF ensemble system is particularly advantageous for heavy rainfall forecasts (in this case, rainfall rates 2 inches per day and greater). The boxes for anomalies between the GFS and ensemble mean suggest that the differences between the two are at least marginally significant at this high threshold.

Fig. 7 raises a note of caution about these results. From day to day there is a very large variation in GSS scores, depending on the rainfall regime dominating at the time. Aggregating over rainfall regimes rather than arbitrarily over sequential time periods thus appears a preferable approach to reach useful verification comparisons.

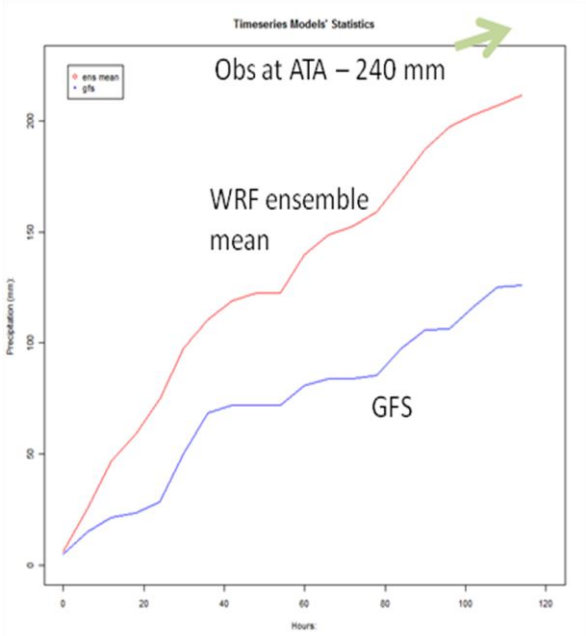

Fig. 4. As in Fig. 3 except for the WRF ensemble mean and for the GFS forecasts. Rain gage observation at Alta is also shown.

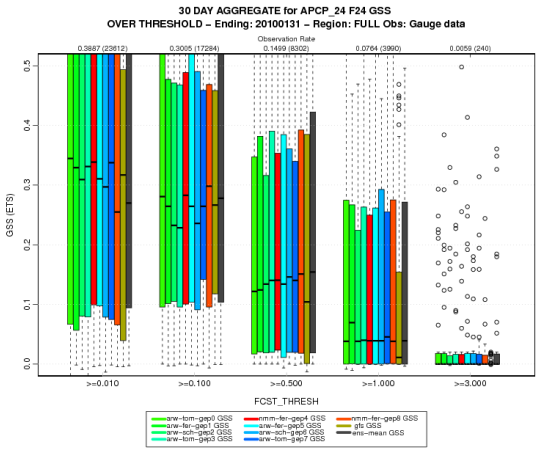

Fig. 5. Distributions of Gilbert Skill Scores (GSS) at five rainfall thresholds for January 2010 for the full WRF domain. Colored boxes are inter-quartile ranges, and the horizontal line for each is the mean of the GSS scores. Verification was performed using daily rainfall totals at gage sites, and forecasts verified are as indicated in legend. Note that the brown and black boxes are GFS and ensemble mean forecasts respectively.

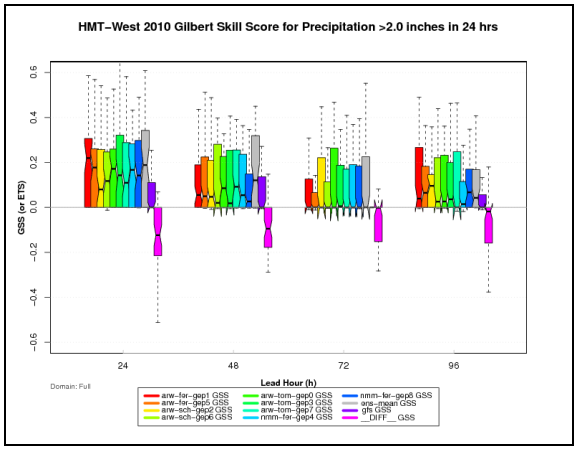

Fig. 6. As in Fig. 5 except for GSS values at a threshold of 2 inches computed at the lead times indicated. The verification period was the full experimental duration (December to March), and score confidence intervals are estimated by the box notches The right-most lavender boxes of each set are the differences between scores for the ensemble mean and the GFS.

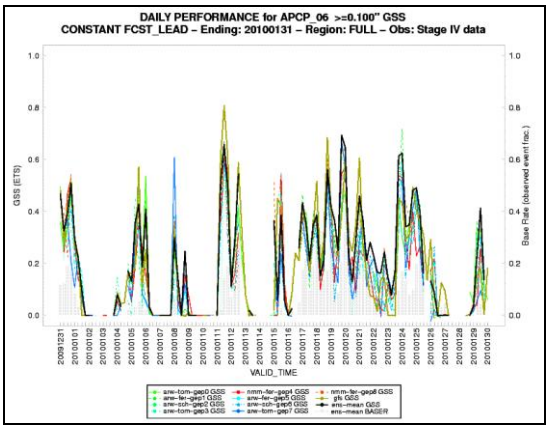

Fig. 7. Time series of Gilberty Skill Score (GSS) during January 2010. Ensemble members, ensemble mean, and GFS forecast are verified against 6h Stage IV gridded precipitation observations.

4. Verification Dataset Impacts

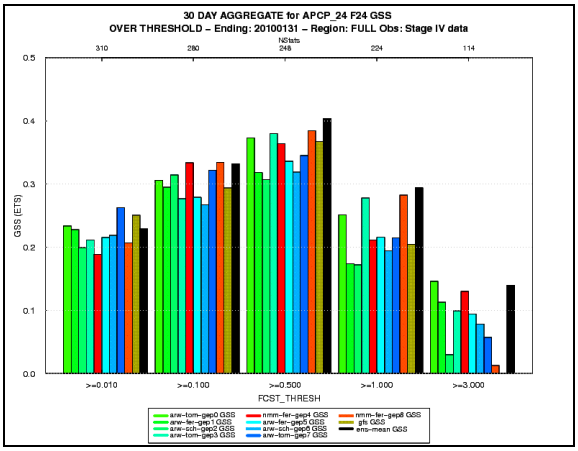

Fig. 8. As in Fig. 5 except for bar plots of GSS verified against 24h Stage IV precipitation analyses.

What impacts can the application of different verification datasets have in a real-life setting? During the winter exercise, verification was computed using both gridded analysis (6h and 24 increments) and point gage observations (24h increments). This offered the opportunity to systematically study verification differences that are due solely to choices of verification data sets. Several factors may be relevant to this difference: rainfall during 6h

accumulation periods cannot reach given thresholds as easily, reducing sampling and negatively affecting ETS scores; and gages are predominantly located in California as opposed to Nevada whereas Stage IV analyses extend across the full domain (excluding Pacific Ocean grid points of course), resulting in verification in poorly-observed geographic regions. Figures 8 and 9, in combination with Fig. 5, reveal a real effect of data sets that impact the interpretation of scoring results. The relative maximum of GSS scores for moderate rainfall thresholds in Fig. 8 is anomalous in the sense that the scores would normally be expected to decrease monotonically with larger thresholds; for verification with gages, for instance, this is indeed the pattern (Fig. 5). The source of this difference appears to be the penalty factor built into the gilbert Skill Score (also called the Equitable Threat Score in deference to this penalty) for situations where there is strong random likelihood of correct forecasts of rainfall. In Nevada, where the Stage IV gridded analyses produce a very large area of light precipitation, this affect is strong at small thresholds. The gage verification, on the other hand, has few gauges in Nevada and thus is less susceptible to this effect. Partial confirmation of this interpretation is provided in the plots of CSI in Fig. 9, which are basically threat scores without this penalty; there, the scores as expected do decrease monotonically with threshold.

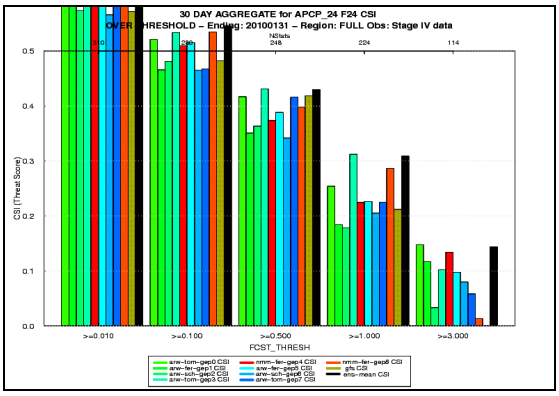

Fig. 9. As in Fig. 8 except for Conditional Success Index (CSI) scores for ensemble member QPF, ensemble mean, and GFS forecast precipitation. Full domain forecasts are verified against 24h gage observations.

5. Conclusions and Further Research

The extensive verification results obtained during the winter experiment in California represent a rich source for studies like those briefly introduced here. In addition to dataset options, the real-time and retrospective scores also offer opportunities for comparing verification within different regions and over various meteorological scenarios.

Acknowledgments

The USWRP provided funding for the website development and other aspects of the research results presented here.