7.2 QUANTITATIVE PRECIPITATION FORECAST (QPF) VERIFICATION OF DWFE

Meral Demirtas[1,2,*], Louisa Nance[1,2], Ligia Bernardet[2,3,4], Ying Lin[5], Andrew Loughe[2,3,6], Jennifer Luppens Mahoney[2,3], Robert Gall[1,2], and Steven Koch[2,3]

[1]National Center for Atmospheric Research, Boulder, CO
[2]Affiliated with the Developmental Testbed Center, Boulder, CO
[3]NOAA Research – Forecast Systems Laboratory, Boulder, CO
[4]Systems Research Group, Inc., Colorado Springs, CO
[5]National Centers for Environmental Prediction, Camp Springs, MD
[6]Cooperative Institute for Research in Environmental Studies, University of Colorado, Boulder, CO

1. INTRODUCTION

Assessing the quality of quantitative precipitation forecasts from numerical weather prediction models has been accomplished by either comparing a precipitation forecast grid to observation stations or to an analysed gridded field of precipitation.

If the aim for evaluating precipitation forecasts is to asses the accuracy of the forecast at certain location, then one needs to interpolate the forecast to the observation location and use the actual observations as verification data (grid-to-point approach). This verification approach is used operationally at the European Centre for Medium Range Weather Forecasts, Deutscher Wetterdienst, Turkish Meteorological Service and some other places. This approach does not smooth observations. Its disadvantage is that it smoothes the forecasted precipitation field, which increases the minima and reduces the maxima. This method does not conserve the total precipitation forecast by the model.

If the aim is to examine multiple forecast models, then one should compare, on a common verification grid, an analysed field of precipitation to the gridded precipitation forecast (grid-to-grid approach). This verification technique is currently used operationally at the National Centers for Environmental Predictions (NCEP) and at the Australian Bureau of Meteorology. Some advantages of this technique are that the gridded (re-mapped) observations better represent the grid-scale quantities predicted by the model, and the sampling is spatially uniform. The technique preserves, for some degree of accuracy, the total forecast precipitation of the native grid, although it introduces some minimal smoothing due to interpolation. Its disadvantage is that the analysis tends to smooth the observations.

* *Corresponding author address:* Meral Demirtas, NCAR/DTC, P. O. Box 3000, Boulder, CO 80307. E-mail: demirtas@ucar.edu

This paper outlines the basic properties of the two QPF verification techniques that employ the two basic approaches introduced above and presents statistical results generated during the DTC Winter Forecast Experiment (DWFE), which was performed from 15th January to 31st March 2005 (For details of DWFE see Nance et al. 2005 and Bernardet et al. 2005a).

2. QPF VERIFICATION TECHNIQUES USED

Verification is critical to the Developmental Testbed Center's (DTC) testing and evaluation process. (For more details about the DTC, see Nance et al. 2005.) The current verification system used at the DTC is explained in detail in Demirtas et al. (2005). The Quantitative Precipitation Forecast (QPF) Verification (QPFV) portion of this system is outlined below.

2.1 QPF Verification-1 (QPF-V1): Grid-to-Grid

For QPF-V1, obtained from the NCEP, all the model precipitation forecasts and available corresponding observations are re-mapped to the same verification grid. The remapping technique involves subdividing the boxes centered on each post-processing grid point into 5x5 sub-grid boxes, and assigning to each sub-grid point the value of the nearest native grid point. The average of these 25 sub-grid point values produces the remapped value of the post-processing verification grid point (Accadia et al. 2003).

The NCEP/CPC's 1/8 degree daily precipitation analysis data set (accumulated 12 UTC-12 UTC) was the verification dataset used for the grid-to-grid verification procedure. This analysis dataset is based upon the 7,000-8,000 daily gauge reports.

2.2 QPF Verification-2 (QPF-V2): Grid-to-Point

For QPF-V2, the Real Time Verification System (RTVS; Mahoney et al 2002), developed at NOAA

Forecast System Laboratory (FSL), was used. Model forecasts were bilinearly interpolated to approximately 4,500 Hydrometeorological Automated Data System (HADS) gauge observation sites. Accumulation periods were 3, 6, 12, 24, 36 and 48 hours, where the starting time of the accumulation period corresponds to the initial time of the model forecast (Loughe et al. 2001).

3. RESULTS FOR DWFE

3.1 Brief Description of DWFE

During DWFE, CONUS-wide, forecasts were generated daily (00 UTC run cycle) using two dynamic cores of the Weather Research Forecast (WRF) framework: the Nonhydrostatic Mesoscale Model (WRF-NMM), developed by NCEP (Janjic, 2003), was run on NOAA/FSL's supercomputer and the Advanced Research WRF (ARW), developed by NCAR (Skamarock, 2005), was run on NCAR's supercomputer. Forecasts for both WRF cores extended to 48-h with output available at 3-h intervals. (For details of DWFE see Nance et al. 2005 and Bernardet et al. 2005a). The WRF-ARW and WRF-NMM were both run with 5-km horizontal grid spacing on their native grid projections.

3.2 Processing Data for QPF-V1 and QPF-V2

QPF verification is highly sensitive to scaling, and different models are best compared at the same resolution, for the QPF-V1 forecasts and observations were remapped on the 5-km grid (referred to as G163) for the two WRF cores and on the 12-km grid (referred to as G218) for the WRF cores and the Eta.

For both QPF-V1 and QPF-V2 systems, dichotomous statistics were computed to verify precipitation at specified thresholds of 0.01, 0.10, 0.25, 0.50, 0.75, 1.0, 1.5, 2.0, 3.0 inches for 24h; the verification domains consist of: the CONUS, the Eastern US, Central US and Western US regional sub-domains. QPF verification statistics were computed for 24-h.

It is important to note that 24-h precipitation accumulation times are different for the two QPF verification techniques. For QPF-V1, precipitation forecasts were accumulated from 12 to 36 hours to be in compliance with the NCEP/CPC's daily precipitation analysis, for QPF-V2, precipitation forecasts were accumulated from 00 to 24 hours.

The following verification statistics were computed for both QPF verification techniques (available at http://www-ad.fsl.noaa.gov/fvb/rtvs/wrf/DWFE/): Bias, Conditional Miss Rate, Critical Success Index (also known as Threat Score), Equitable Threat Score, False Alarm Rate, Probability of False Detection, False Alarm Ratio, Heidke Skill, Peirce-Hanssen-Kuipers Skill Score (also referred to as the True Skill Statistic), Probability of Detection and Threat Score. (For definitions of these scores, readers are advised to see Wilks, 1995.)

3.3 A Summary of QPF-V1 Results for DWFE

Overall QPF verification results for the grid-to-grid evaluation of the WRF-ARW and the WRF-NMM indicated that for all domains and, regardless of the verification grid resolution, both WRF-cores overforecast precipitation, while the Eta model tended to under-forecast precipitation.

The ETS skill scores for the Eta model and the WRF-cores were very similar to each other for all domains and for all threshold values. An ETS scores of 1 is a perfect score. The maximum ETS skill score noted in this study was around 0.5 for the low threshold values. For high thresholds, ETS skill scores approached 0.0, indicating no skill.

Comparisons of the QPF-V1 results obtained for the two verification grids (G163 and G218) for all the domains are summarized in sections i, ii, iii, and iv. (The Eta model forecasts have not yet been re-mapped onto G163 yet, therefore only the results for WRF-cores are highlighted.)

i. Highlights for the CONUS domain:

The difference between the ETS skill scores for G218 and G163 were small for all threshold values (Fig. 3.1). All models had very low skill scores for high threshold values.

Comparisons of the bias scores obtained on G163 and on G218 indicated the differences were very small (Fig. 3.2). Results of both verification grids showed that the WRF-cores over-predicted the precipitation for the high threshold values, while the Eta model under-predicted.

ii. Highlights for the Eastern domain:

A comparison of the ETS skill scores showed that both WRF cores had slightly higher ETS skill scores on G218 (Fig. 3.3) except at 0.01 inch threshold. The highest ETS skill score was achieved around 0.5 for both WRF-cores for the low threshold values.

The bias scores obtained on G218 and on G163 were similar for the threshold values smaller than 2.0 inches (Fig. 3.4). For the threshold values greater than 2.0 inches, bias scores obtained on G163 were larger these obtained on G218. As it was noted for the CONUS, the WRF-cores had a tendency to over-predict the precipitation; this over-prediction was more pronounced at the high threshold

iii. Highlights for the Central domain:

ETS skill scores for the Central domain indicate the performance of the models were similar. The WRF-ARW had slightly higher scores on G163 compared to the results obtained for G218 (Fig. 3.5). Overall, the skill decreased as the threshold values increased. Over the Central domain, the WRF-cores had the least skill compared to the other domains.

The bias scores obtained on G163 and on G218 were noticeably different for the threshold values greater than 1.0 inch (Fig. 3.6). Bias scores obtained on G163 were larger than those obtained on G218. The over-prediction of the precipitation at high threshold values was evident for the Central domain.

iv. Highlights for the Western domain:

A comparison of the ETS skill scores showed that the scores were slightly higher on G218 compared to the results obtained for G163 (Fig. 3.7). This difference was particularly noticeable for the WRF-NMM. The ETS skill score is sensitive to the grid transformation process. Relatively small changes in hits, misses, and false alarms affect ETS, particularly at higher threshold values, where the number of correct no-rain forecasts is much larger.

The bias scores obtained on G218 and on G163 (Fig. 3.8) were different for the threshold values larger than 1.5 inches. The WRF-ARW bias scores obtained on G163 were smaller than those obtained on G218, while the opposite was noted for the WRF-NMM.

3.4 A Comparison of QPF-V1 and QPF-V2 Results for the CONUS Domain

The results of QPF-V1 obtained on G218 and QPF-V2 were compared for the 24h accumulation period. For this study the same precipitation accumulation time was used. This comparison including the Eta model indicated both verification techniques lead to similar conclusions (Fig. 3.9 and 3.10). Differences in ETS skill scores between the Eta and the two WRF models were insignificant for all rainfall thresholds (Figure 3.11 and 3.12). This was particularly noticeable for threshold values smaller than 1.50 inches. The bias scores (for threshold values larger than 1.50 inches), obtained for QPF-V1 (Figure 3.13 and 3.14), were slightly higher than the results of QPF-V2. This difference was expected because the verification approaches and the observations used to evaluate the forecasts were different.

Regardless of the QPF verification technique used, both WRF-cores showed a distinct tendency to over-predict precipitation at high threshold values, while the Eta model had the opposite tendency, under-predicted precipitation at the high thresholds.

4. DISCUSSION AND CONCLUSIONS

Looking in detail at the results obtained from QPF-V1, a comparison of G163 and G218 results for 24-h precipitation accumulation period implied that for the CONUS, the differences in the statistical results were small. The scene was different for the regional domains. The remapping technique preserves the total rain amount. Since remapping to a coarser grid yields some smoothing, as a consequence, the area covered by the low threshold values increases and the area covered by the high threshold values decreases. Therefore, it did not come as a surprise that bias scores obtained for G163 were greater for the high threshold values than G218.

A comparison of QPF-V1 (for G218) and QPF-V2 results for 24-h accumulated precipitation indicated that both verification techniques lead to similar conclusions for the CONUS. Some differences were noted for the high threshold values. These differences could be attributable to the different observational data sets and verification techniques used.

The results clearly indicated that both WRF-cores had the lowest QPF skill over the Central domain when compared to the Eastern or the Western domains.

An error was discovered in the DWFE version of the WRF-NMM in its radiation parameterization. This error impacted the interactions between the short and the long wave radiations and ice clouds. As a consequence, excessive solar radiation arrived at the surface during the day, while there was a deficit in long-wave radiation loss to space at night (Bernardet et al 2005). The impact of this error on the WRF-NMM QPF performance is unknown at this point.

In QPF-V1, the remapping, by its construction, gives a reduced edge smoothing on precipitation forecasts, while the precipitation maxima are not changed very

much by the simple average. The technique preserves the total precipitation amount. Therefore, the skill scores are less affected by smoothing introduced by the remapping. On the other hand, relatively small changes in hits, misses, and false alarms affect ETS, particularly at higher threshold values, where the number of correct no-rain forecasts is much larger.

In QPF-V2, the smoothing effect of bilinear interpolation on the forecast precipitation field creates a decrease in the original maxima and an increase in the original minima. In return this decreases the bias scores. The smoothing also produces a smooth field, while decreasing the gradients across the rain and no-rain boundaries. If precipitation above a certain threshold is not observed, this edge-smoothing effect, introduced by the interpolation, may decrease the forecast precipitation in such a way that a false alarm becomes a correct no-rain forecast. Since ETS is sensitive to the hits, this affects the ETS scores.

The evaluation of high resolution model products, using standard verification techniques is not sufficient. Classic verification techniques give basic information about the performance of spatial forecasts. They are not diagnostic and they may not give information needed to improve the forecasts. The development of new verification techniques is an active area of research. Advanced techniques may help to quantify errors in occurrence, location, magnitude, size and shape. New verification techniques must be employed such as the entity-based, object-oriented (Bernardet et al. 2005b), and scale decomposition techniques currently under development by the research community. Some of these new techniques will be added to the DTC's verification system when their capabilities have been demonstrated.

Acknowledgements: We would like to thank Ms. Tressa Fowler (NCAR/RAL) for computing confidence intervals. Suggestions received from Dr Barbara Brown (NCAR/RAL) were very useful.

5. REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbor Average Method on High-Resolution Verification Grids, *Wea. Forecasting*, **18**, 918–932.

Bernardet, R. L., L. Nance, H. Chuang, A. Loughe, M. Demirtas, S. Koch and R. Gall, 2005a. The Developmental Testbed Center Winter Forecasting Experiment. *21st Conference on Weather Analysis and Forecasting*, 1-5 August, Washington, D.C., American Meteorological Society.

Bernardet, R. L., P. Bogenschutz, J. Snook and A. Loughe, 2005b: WRF forecasts over the Southeast United States: Does a larger domain lead to better results? *6th WRF/15th MM5 Users' Workshop*, 27-30 June 2005, Boulder, Colorado.

Demirtas, M., L. Nance, L. Bernardet, Y. Lin, H.-Y. Chuang, A. Loughe, R. Gall, and S. Koch, 2005: Verification Systems used for the Developmental Testbed Center. *6th WRF/15th MM5 Users' Workshop*, 27-30 June 2005, Boulder, Colorado.

Janjic, Z. I., 2003. A nonhydrostatic model based on a new approach. *Meteorol. Atmos. Phys*. **82**, 271-285.

Loughe, A. F., J. K. Henderson, J. L. Mahoney, and E. I. Tollerud, 2001: A Verification approach suitable for assessing the quality of model-based precipitation forecasts during extreme precipitation events. Preprint, *Symposium on Precipitation Extremes: Prediction, Impacts, and Responses*, 14-19 January 2001, Albuquerque, NM. American Meteorological Society.

Mahoney, Jennifer Luppens, Judy K. Henderson, Barbara G. Brown, Joan E. Hart, Andrew Loughe, Christopher Fischer, and Beth Sigren, 2002: The Real-Time Verification System (RTVS) and its Application to Aviation Weather Forecast. *10th Conference on Aviation, Range, and Aerospace Meteorology*, 13-16 May, Portland, OR.

Nance, L., L. Bernardet, H-Y. Chuang, G. DiMego, M. Demirtas, R. Gall, S. Koch, Y. Lin, A. Loughe, J. Mahoney, and M. Pyle, 2005: The WRF Developmental Testbed Center: *6th WRF/15th MM5 Users' Workshop*, 27-30 June 2005, Boulder, Colorado.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang and J. G. Powers, 2005: A Description of the Advanced Research WRF Version 2, *NCAR Tech Note*, NCAR/TN–468+STR, 88 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO, 80307]. Available on-line at: http://box.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf)

Wilks, D. S., 1995: *Statistical Methods in Atmospheric Sciences.* Academic Press, 467 pp.

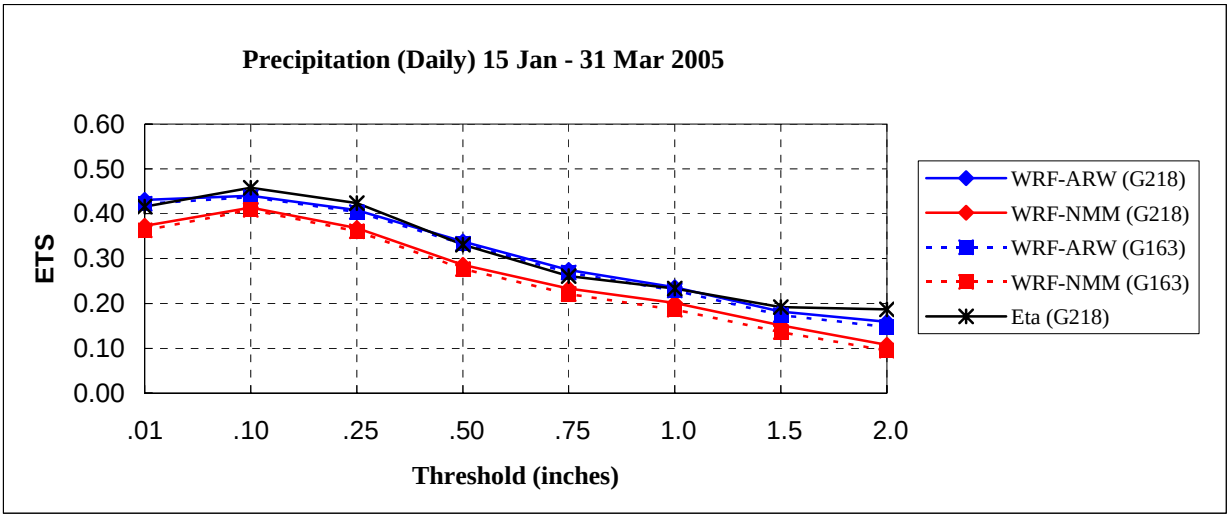

Figure 3.1 QPF-V1 ETS results for 24h precipitation accumulation period for the CONUS domain.

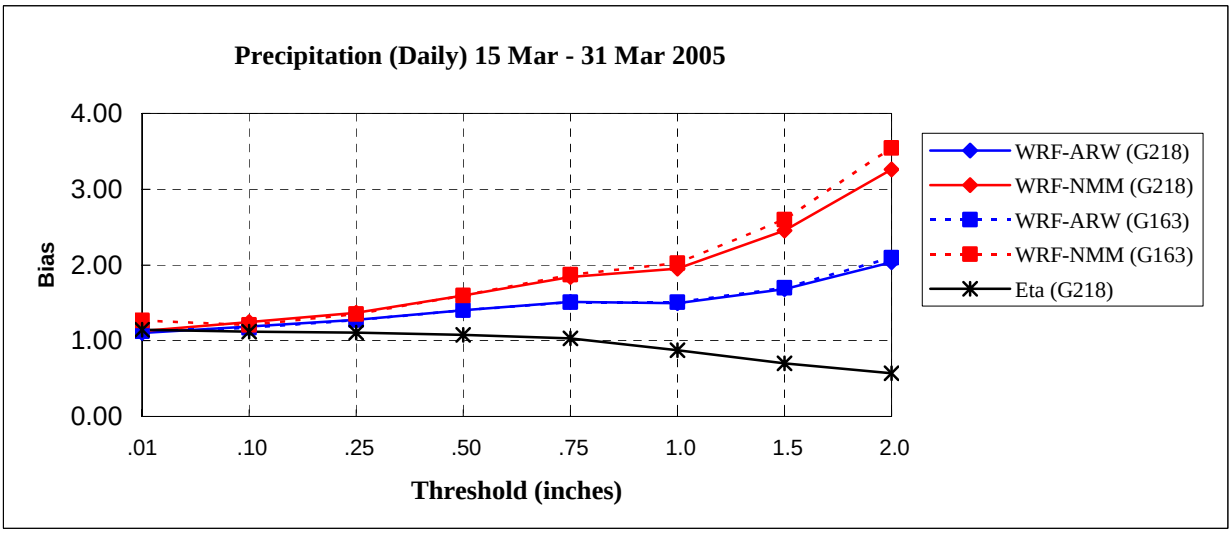

Figure 3.2 QPF-V1 bias results for 24h precipitation accumulation period for the CONUS domain.

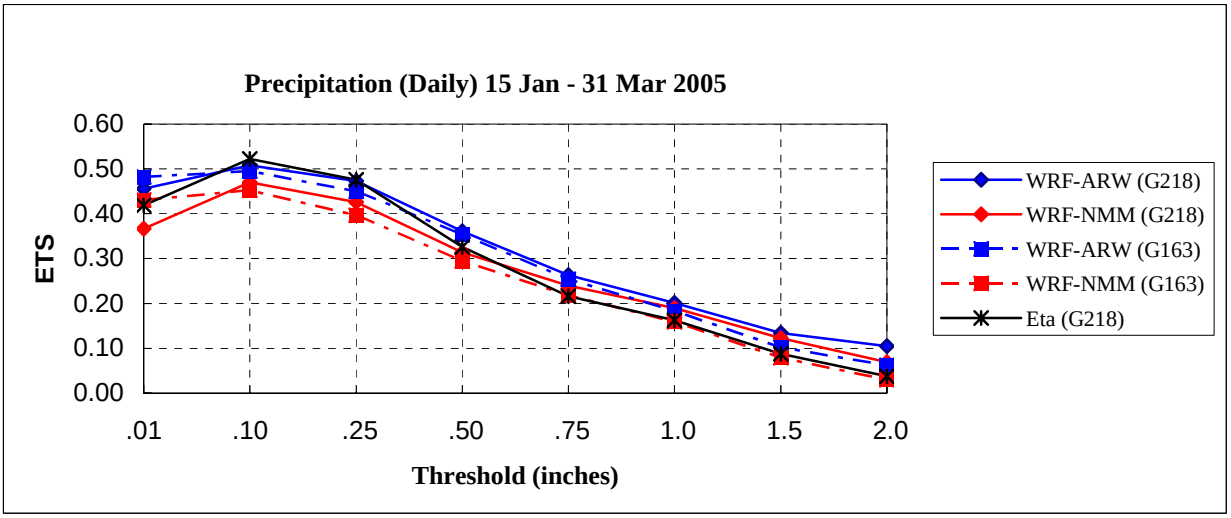

Figure 3.3 QPF-V1 ETS results for 24h precipitation accumulation period for the Eastern domain.

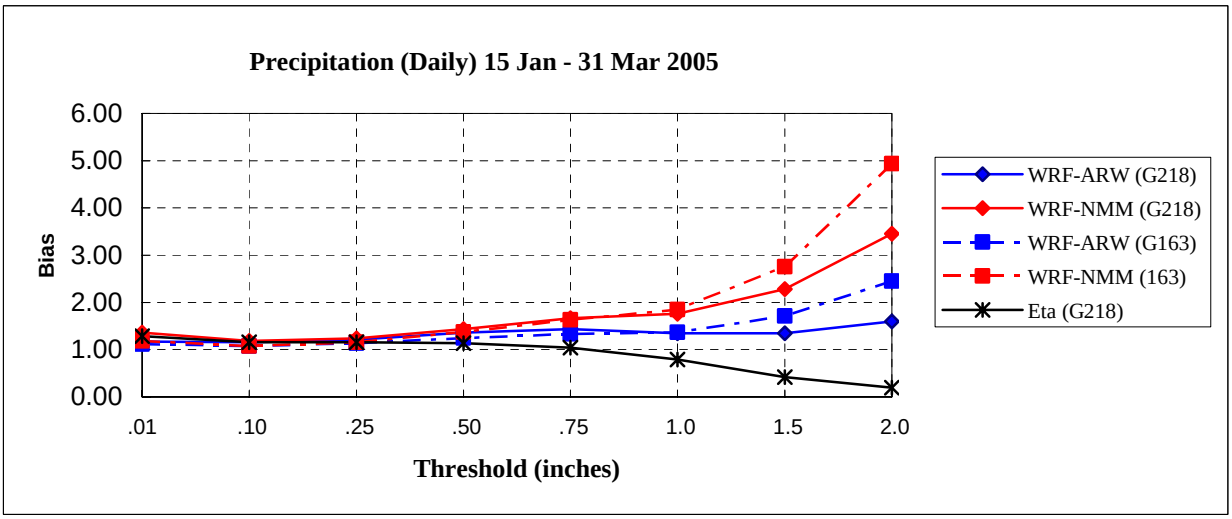

Figure 3.4 QPF-V1 bias results for 24h precipitation accumulation period for the Eastern domain.

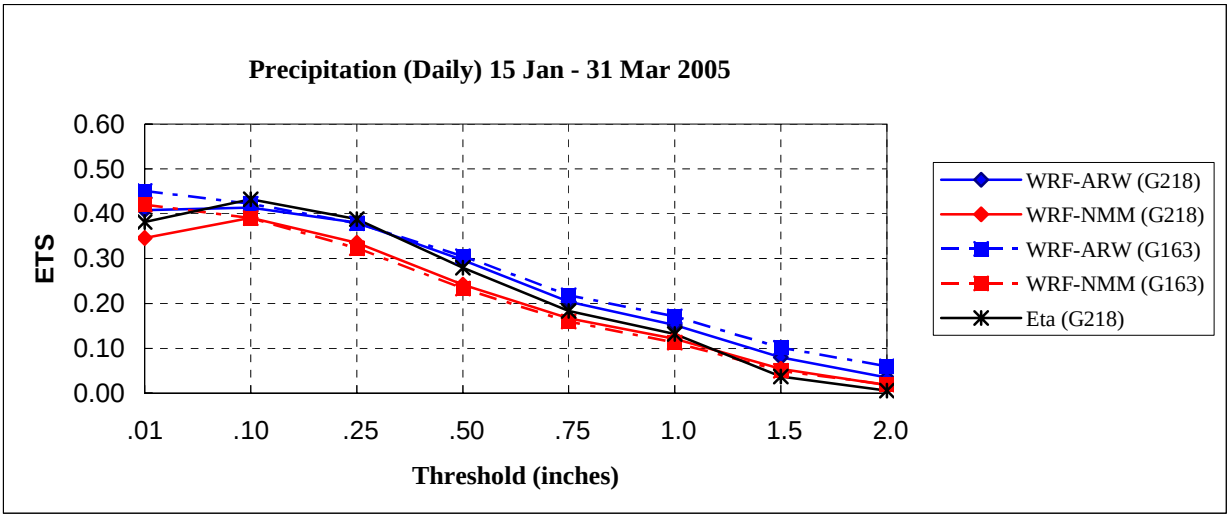

Figure 3.5 QPF-V1 ETS results for 24h precipitation accumulation period for the Central domain.

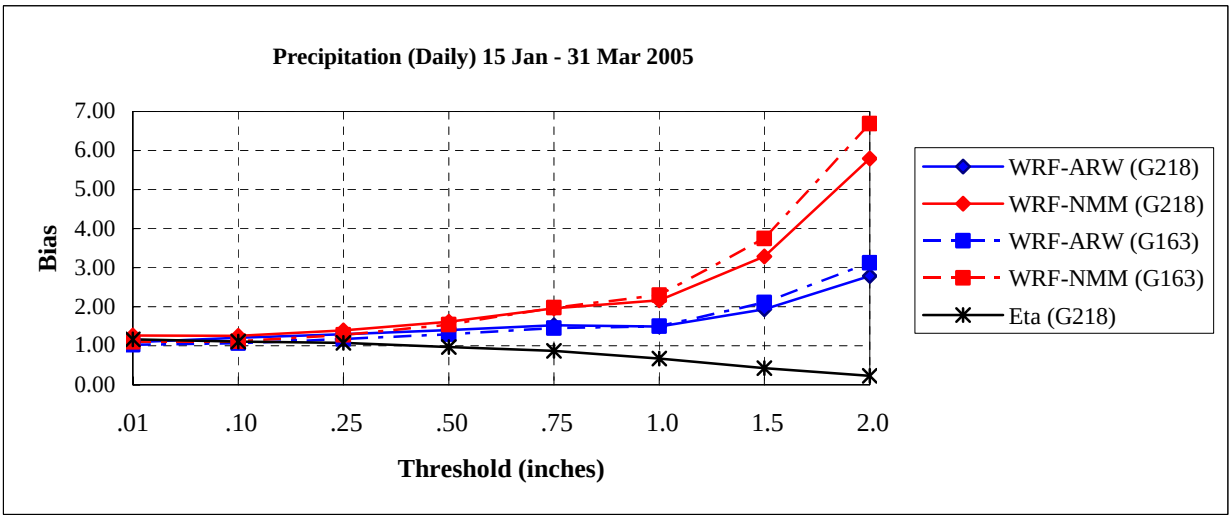

Figure 3.6 QPF-V1 bias results for 24h precipitation accumulation period for the Central domain.

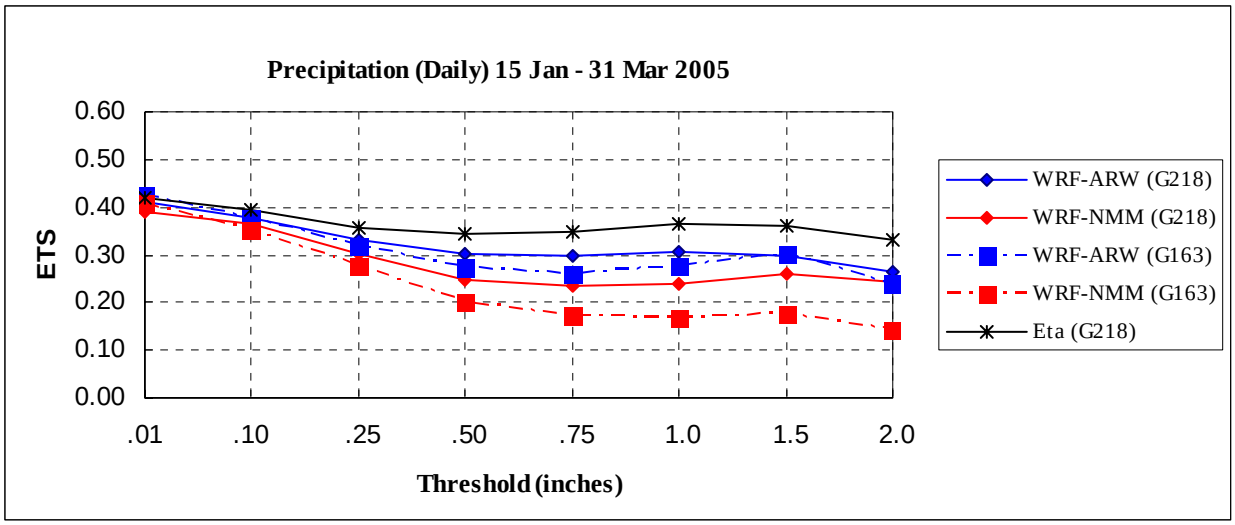

Figure 3.7 QPF-V1 ETS results for 24h precipitation accumulation period for the Western domain.

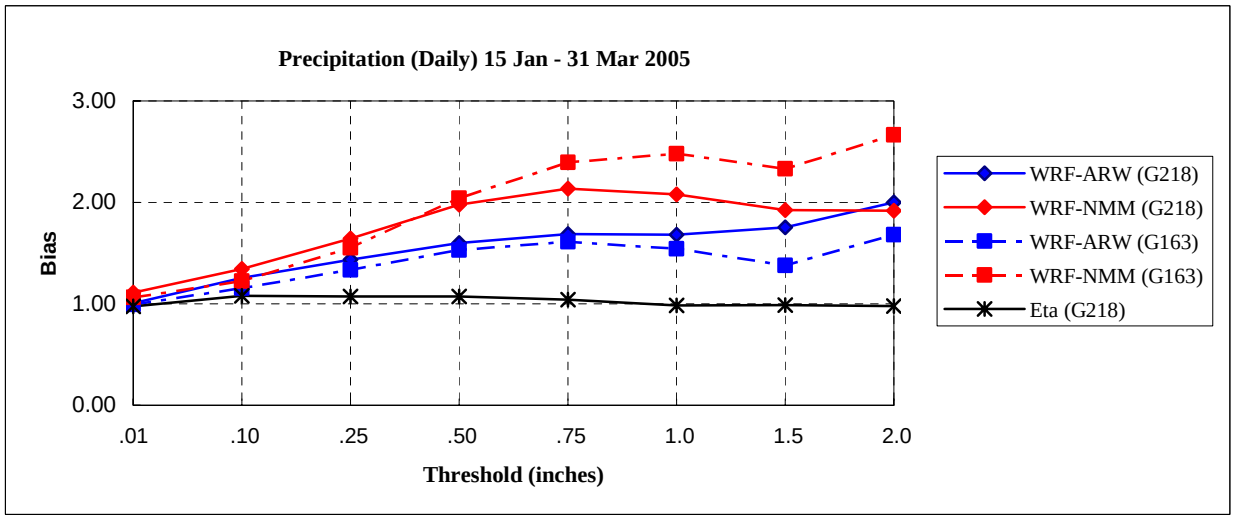

Figure 3.8 QPF-V1 bias results for 24h precipitation accumulation period for the Western domain.

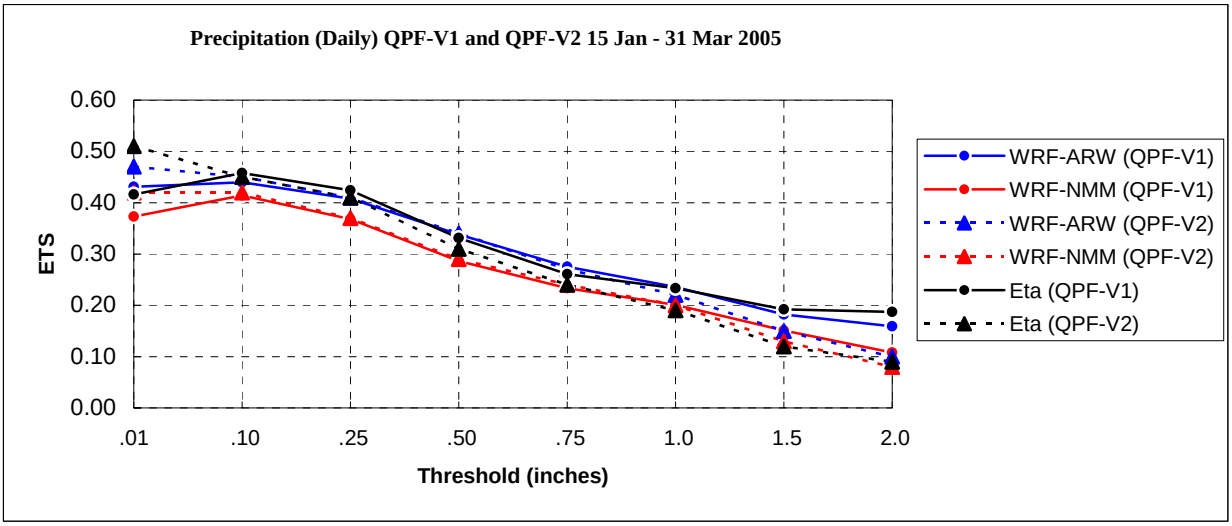

Figure 3.9: Bias scores obtained from QPFV-1 (on the G218) and QPFV-2 techniques.

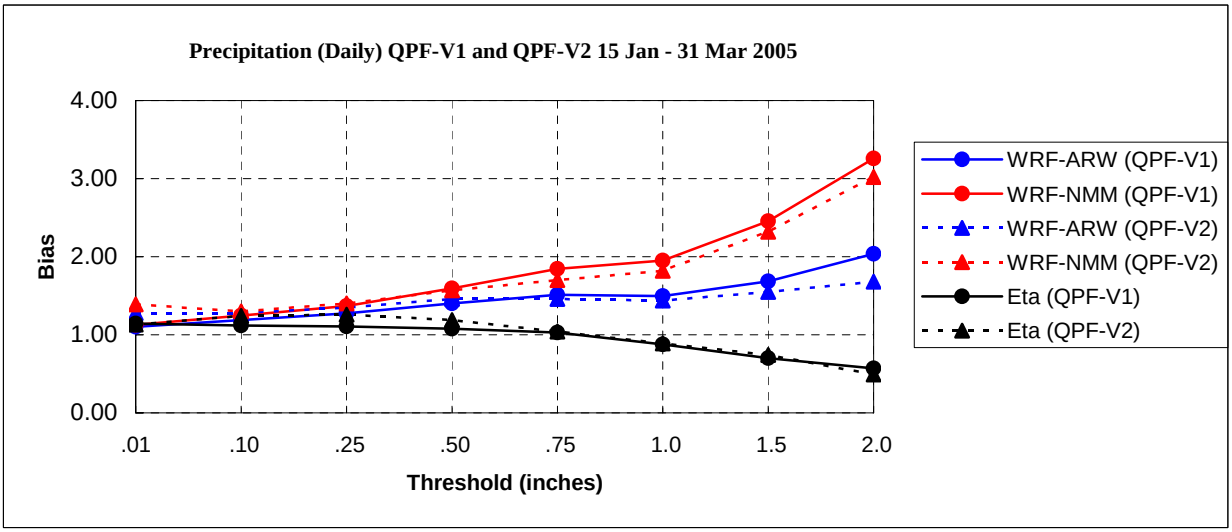

Figure 3.10 ETS results obtained from QPFV-1 (on the G218) and QPFV-2 techniques.

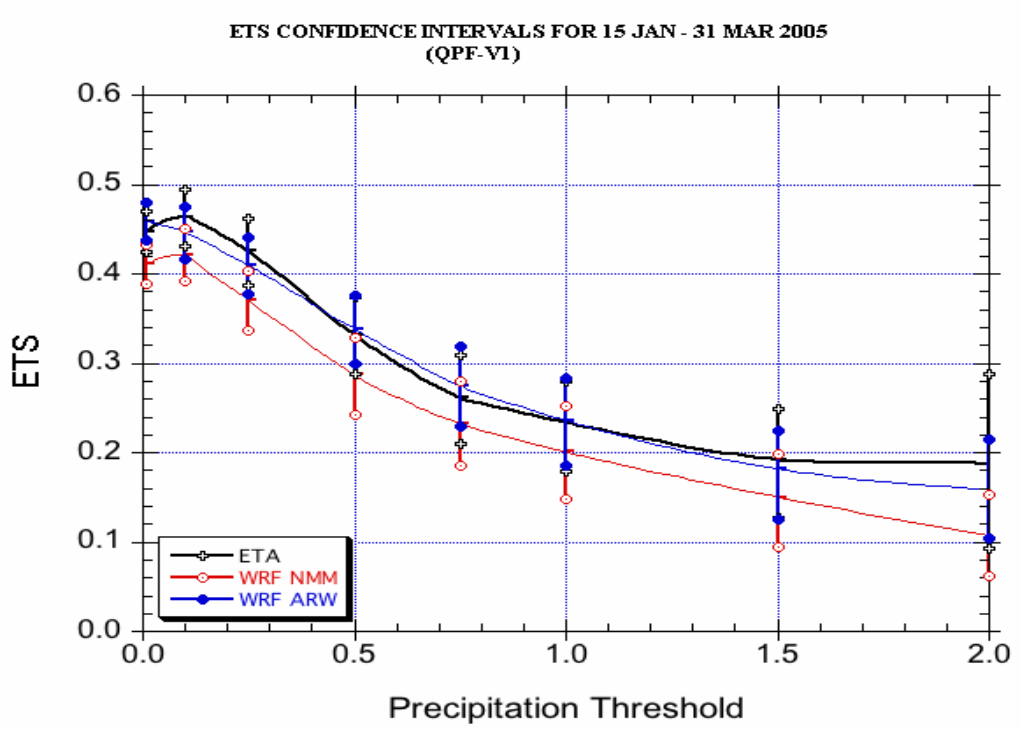

Figure 3.11 QPF-V1 (on G218) ETS results for 24h precipitation accumulation, 95% confidence intervals using bootstrapping method are shown.

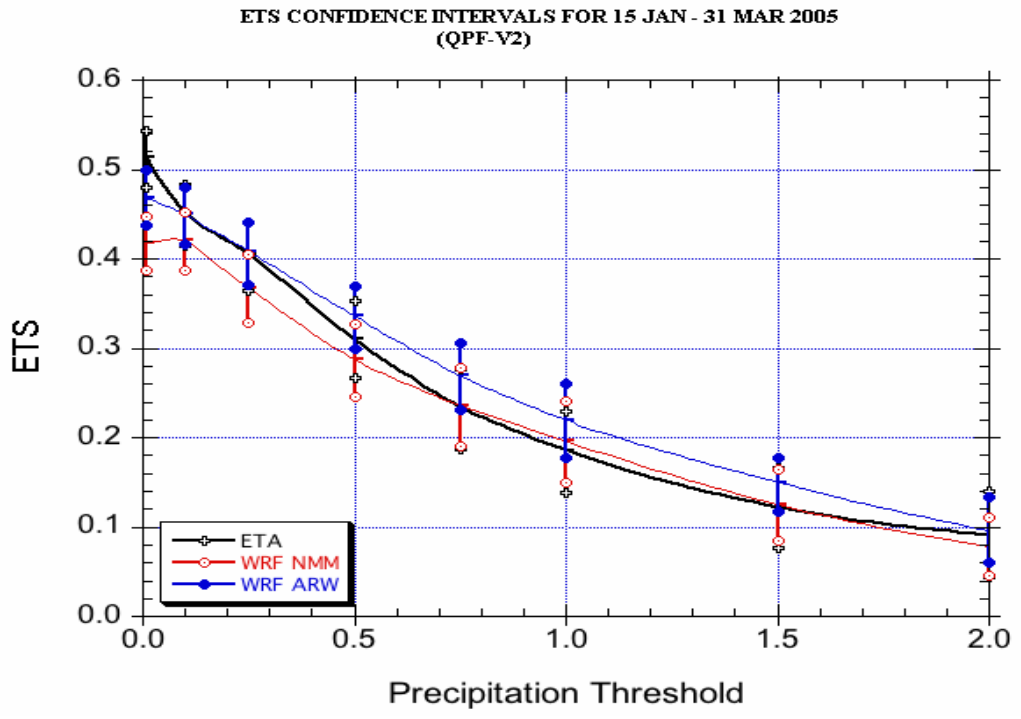

Figure 3.12 QPF-V2 ETS results for 24h precipitation accumulation 95% confidence intervals using bootstrapping method are shown.

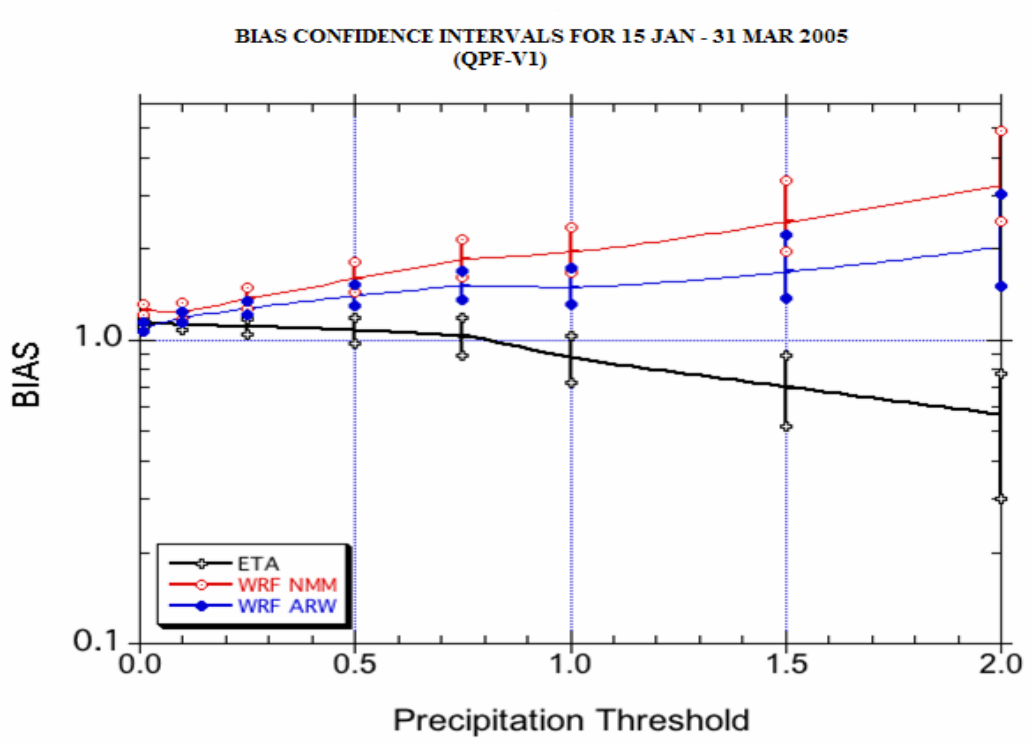

Figure 3.13 QPF-V1(on G218) bias results for 24h precipitation accumulation 95% confidence intervals using bootstrapping method are shown.

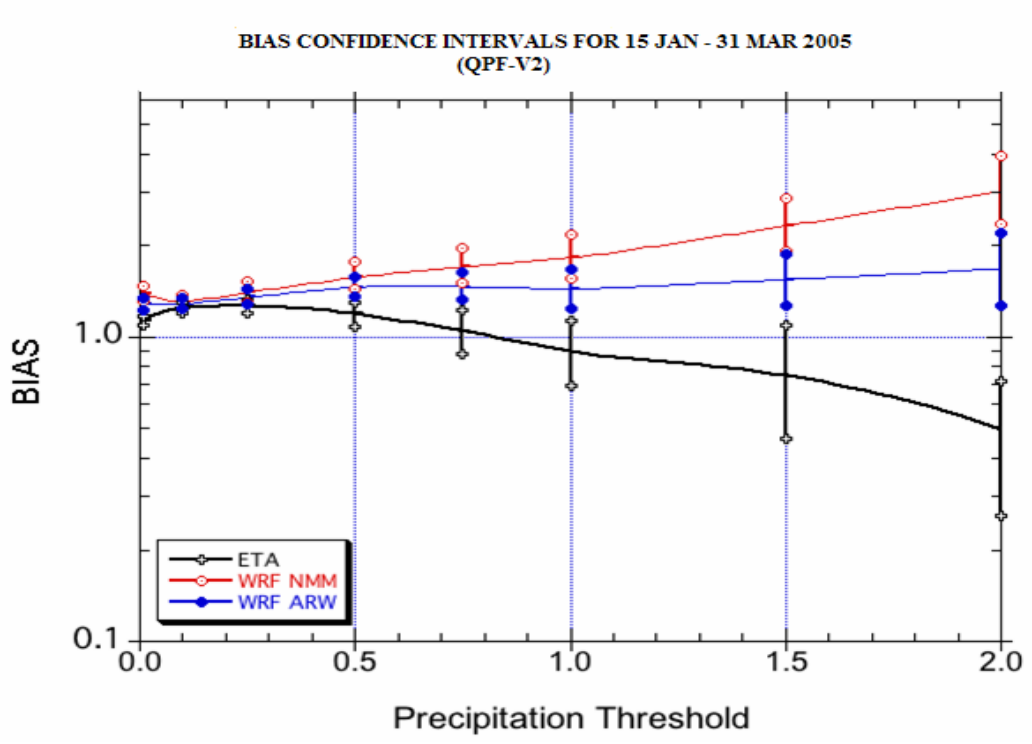

Figure 3.14 QPF-V2) bias results for 24h precipitation accumulation 95% confidence intervals using bootstrapping method are shown.