Reviewing the SPC/NSSL Spring Program 2003: An Evaluation of the Use of Short-Range Ensemble Forecasting Systems and New High Resolution Deterministic Models in the Prediction of Severe Thunderstorms.

Marc R. Dahmer
University of Missouri, Columbia, Missouri

Steven J. Weiss
David R. Bright
Storm Prediction Center, Norman, Oklahoma

Jack S. Kain
National Severe Storms Laboratory, Norman, Oklahoma

NOAA/NWS/Oak Ridge Institute for Science and Education
Final Project

01 August 2003

Corresponding Author:
Marc R. Dahmer
Atmospheric Science
University of Missouri-Columbia
373 McReynolds Hall
Columbia, MO 65211

Email: mrdad4@mizzou.edu

Abstract

The SPC/NSSL Spring Program is typically held during the heart of the severe convective weather season in Norman, Oklahoma. This is an opportunity for researchers and operational meteorologists to interact and collaborate on a variety of experimental forecast and other operationally relevant research programs. This year's program focus was two-fold. The primary objectives were to explore the use of Short-Range Ensemble Forecasting (SREF) systems to provide meaningful guidance in severe weather forecasting, and to examine the ability of new high-resolution deterministic models to predict convective initiation and evolution. These objectives were subjectively analyzed by participants and evaluated for its operational forecasting uses. The participants of the program were also surveyed to glean insight into the program's utility. Through the evaluation of the objectives, it was found that the SREF output does have positive use operationally. It was also found that just because a model's QPF is initially misplaced or missing, does not mean the model should be discounted as a tool in the prediction of severe weather as it pertains to watch lead time.

1.) Introduction

Predicting severe convective weather is a subject that remains at the forefront of research. Leading the way, in this respect, is an acute responsibility at the Storm Prediction Center (SPC) and the National Severe Storms Laboratory (NSSL) in Norman, Oklahoma. The co-location of these two agencies has resulted in much collaboration on operationally applicable research programs. The goal of these programs is to provide forecasters with better tools and knowledge to improve severe weather forecasting. During the spring of 2000 and 2001, these collaborative programs focused on critical SPC operational products including the short-term predictability of severe and non-severe thunderstorms and potential impact on operational convective watch lead-time. During the spring of 2002, the program focused on providing forecast support for the IHOP field project, primarily addressing afternoon convective initiation and nocturnal MCS development. The goal of the Spring Program, every year, is to foster interaction between SPC forecasters, NSSL scientists, educators, and other researchers and forecasters that will lead to improved knowledge in the arena of severe weather prediction.

During Spring Program 2003, the primary objectives were to: 1) Explore the utility of Short-Range Ensemble Forecasting (SREF) systems to provide unique and meaningful guidance in operational severe weather forecasting, and 2) Examine the ability of new high-resolution models to predict convective initiation and evolution, as it relates to improving watch lead time. This was accomplished by assembling researchers, forecasters, and educators from across the nation as well as overseas. The program was

conducted over eight weeks from April 14th through June 6th. Full-time participants worked the entire weekday while part-time visitors worked two to three days out of the week. Each week had a different team whose main goal was to subjectively evaluate the SREF system and for new products or ideas that would assist forecasters in more accurately predicting severe convective weather, and evaluate deterministic models for their ability to predict convective initiation.

2.) Short Range Ensemble Forecasting (SREF) Systems

a) Data and Methodology

The use of ensemble methodologies has resulted in dramatic improvements in the skill of medium-range weather forecasts (Tracton and Kalnay 1993; Toth and Kalnay 1993; Molteni et al. 1996). The 2003 Spring Program was designed to test usefulness of ensemble forecasting as a supplementary tool in short-range severe weather forecasting. An initial Day 2 outlook was created using 1200 UTC deterministic model output. The creation of a Day 2 outlook was chosen because it depends almost entirely on model output, while a Day 1 outlook would involve the use of observational data as well as model output. After the initial Day 2 outlook is produced, input for the creation of the MM5 SREF perturbations, based on what the forecast team deemed 'problem areas' in the forecast, were submitted. The MM5 and NCEP SREF output were the only tool utilized to modify the preliminary outlook and create the final Day 2 outlook. These outlooks are verified by severe storm reports collected from local storm reports (LSR) issued by NWS WFOs across the country. By taking this approach, the two outlooks could be compared and the impact of using the SREF output can be assessed. One

evaluation metric for the rating of the initial outlook was performed by the participants' subjective analysis of how well the outlook probabilistic regions captured LSRs in space and concentration. The outlook is rated on a scale of zero to ten, with 'zero' being a poor forecast and 'ten' being a nearly perfect forecast. The final Day 2 outlook was rated in comparison to the initial outlook to assess if the level of accuracy changed. The difference of the ratings for each day was assessed to see how often the SREF output assisted in improving the initial outlook. A Brier Score (BS) and Relative Operating Characteristic (ROC) were also applied to each forecast. Then a percentage improvement over the initial forecast was calculated for each.

Percent improvement over Initial fcst using BS
$1 - (BS_{final} / BS_{initial})$

Percent improvement over initial fcst using ROC
$(ROC_{final} - ROC_{initial}) / (1 - ROC_{initial})$

In addition to the ratings, comments on why a forecast was rated a certain way and a forecast discussion were included for each outlook. Ratings were also assigned to the different SREF output fields, such as spaghetti charts, mean/spread charts, and probability charts. This was done to evaluate what forecasters were partial to about the SREF output. The SREF data was gathered from the Spring Program 2003 website, located at http://www.spc.noaa.gov/exper/Spring_2003.

b) Results

Conducting the analysis of the Day 2 outlook ratings and SREF data led to many conclusions. First of all, the subjective analysis of the outlooks revealed an improvement of eight points over thirty-one total forecast periods. Fourteen of these days experienced

a positive change to the final outlook, while only six days experienced a negative change. 81% of all final Day 2 outlooks experienced no change or received a higher subjective rating that the initial Day 2 outlook. 45% of all final Day 2 outlooks received a higher subjective rating than the initial Day 2 outlook. The Brier score allowed us to objectively analyze the initial and final outlook performance. Fifteen of these days experienced a positive change to the final outlook with ten days experiencing a percentage increase of over 1%, while twelve days experienced a negative change according to the Brier score. Only four of those twelve days experienced a percentage drop of more than 1%. The Relative Operating Characteristic (ROC) is another tool that allowed us to objectively analysis the initial and final outlook performance. Fourteen of these days experienced a positive change to the final outlook with ten days experiencing a percentage increase of over 2%, while eleven days experienced a negative change according to the ROC. Only five of those eleven days experienced a percentage drop of more than 2%. Upon gathering these percentages, more analysis had to be done to determine what SREF output products the participants felt could be used in improving severe weather forecasting.

After sifting through the participant comments, it was found that the most useful charts from the SREF output were the probability charts, followed by the mean/spread charts. Most of the time these charts were used to focus attention on a particular area or to gain confidence in a previous forecast. Some participants gave the impression that the probability charts appeared to provide guidance in defining regions where favorable ingredients were overlapped. The spaghetti charts were good for identifying flow patterns using 500mb heights, locating drylines using dewpoint fields, and for comparing

the of the Operational Eta to the rest of the ensemble members. The spaghetti charts were also said to be too messy to process and difficult to read. One product that was liked and used often was the Pcpn/CAPE/Shear combo. This product combined convective three-hour precipitation, surface CAPE, and sfc–6 km shear. Most of the time, this product was used to solidify changes or provide confidence to an area of concern. Overall, the participants rated the SREF output as moderately to extremely useful in assessing severe weather potential, which was one of the objectives of the Spring Program.

3.) Deterministic Models

a) Data and Methodology

Deterministic models are one of the current tools used in the prediction of severe convective weather. As more research is done, new high-resolution deterministic models are developed and need to be examined as a resource for predicting convective initiation and evolution. This is one of the objectives the Spring Program 2003 seeks to fulfill. Participants of the Spring Program performed a subjective evaluation of each deterministic model's 3hr accumulated precipitation for two separate periods, 1800-2100 UTC and 2100-0000 UTC. This process will help assess how each model behaves with time. Each forecast focused on the regional domain used by the 3 km WRF model that was determined by where the 1300 UTC SPC Day 1 severe outlook had the greatest potential for severe weather. Verification was achieved by comparing the 3hr accumulated precipitation from the Eta12, EtaKF22, NMM8, WRF12, and WRF3 to 3hr accumulated images of radar base reflectivity. After the comparison was complete, a raw

rating of zero to ten was assessed to each individual model for each individual period with a score of 'ten' being given for an excellent forecast. These raw ratings were then averaged for each individual model by period and all periods together. These results can be ambiguous because the benchmarks used to estimate model performance vary from forecast to forecast and forecaster to forecaster. For example, one perfect forecast might be the prediction of no precipitation, while the next event may require extremely realistic timing and evolution of a complex mesoscale convective structure for perfection. To combat this inconsistency, a relative ranking of the raw scores is created. For example, if for a particular forecast period one model out of four was given a rating of 8, two received 7s, and one received a 4, the relative rankings would be 4, 2.5, 2.5, and 1, respectively (Kain et al. 2002). After the relative ratings were produced, they were averaged for each individual model by period and all periods together. As I stated before, this will give an idea of how well, relative to the other deterministic models, a model predicts convective initiation, structure, and mode over time. Scatterplots were also used as a method to evaluate one deterministic model's output to another. These were produced using the raw ratings for each period. Paired t-test scores were also computed to ascertain the statistical significance of any differences between the deterministic models. A t-test score of 0.05 indicates that differences are significant at a 95% confidence level, and this value is often used as a threshold to distinguish between significance and non-significance (Kain et al. 2002). Although, with only nineteen cases, utilizing this data set makes it difficult to gain any statistical significance from the paired t-test scores. Therefore, the t-test scores will not be discussed in the results, but the figure of the scores will be located at the conclusion of this paper (Fig. 18).

b) Results

At first glance, one would assess the results of the subjective analysis of the deterministic models by stating that the EtaKF seemed to do a better job of predicting convective initiation and mode for both periods. While true, I found that the data could be analyzed for much more than that. When analyzing the 'mean raw score' for all periods combined, the EtaKF was most accurate (with 5.39 out of a 10 point score) in predicting convective initiation, structure, and mode followed by the WRF12 (4.87), Eta12 (4.82), NMM (4.79), and WRF3 (3.55) based on thirty-eight total forecasts (Fig. 1). In contrast, the 'mean rank' for all periods combined show that once again the EtaKF was most accurate (with 3.66 out of a 5 point score) followed by the Eta12 (3.13), NMM (3.09), WRF12 (3.08), and the WRF3 (2.04) with a five point score being the best (Fig. 2). The 'mean rank' was utilized for reasons stated earlier. Most times, when a model is off initially, it is typically discounted as a valid resource for predicting convective weather. I have found through the analysis of the deterministic model results from this program that even though a model's QPF is "off" initially, does not mean that it will not assist in predicting convective weather in later periods. Three of the five deterministic models examined experienced a higher mean raw score and mean rank in the second period (2100-0000 UTC) than in the first period (1800-2100 UTC). The WRF3 experienced the greatest increase in 'mean rank' from one period to the next by 21.8%. That increase was followed by an 11.1% increase in the Eta12 and a 4.11% increase in the NMM from the first period to the next (Fig. 4). This showed that the forecast team, on average, expressed higher confidence in the second period forecast of the WRF3,

Eta12, and NMM. The WRF12 exhibited the steepest drop-off in 'mean rank' from one period to the next with a decrease of 18.6%. That was followed by a decrease in the EtaKF of 9.64% (Fig. 4). On the whole, the analysis of raw scores shows that the highest confidence was expressed in forecasts from the 1800-2100 UTC EtaKF, followed by the 1800-2100 UTC WRF12, the 2100-0000 UTC EtaKF, the 2100-0000 UTC Eta12, the 2100-0000 UTC NMM, and the 1800-2100 UTC Eta12 and NMM (Fig. 3).

Another way to gain insight into the different high-resolution models is by examining scatterplots of each model's raw scores compared to each other. Scatterplots give an idea as to which model performed better at predicting convective initiation, structure, and mode. When analyzing Eta12 and the NMM, not much scatter occurred between the two models, during both periods, suggesting they verified rather equally throughout the program (Fig. 5 & 6). The comparison of the EtaKF and WRF12 exhibited more scatter in both periods than the previous comparison of the Eta and NMM (Fig. 7 & 8). Figure 8 shows that the EtaKF verified better and more often than the WRF12. Having a greater number of plots on EtaKF side of the x-y line demonstrates this. Further examples are provided as figures.

4.) Participant Evaluation

a) Data and Methodology

Participants of the Spring Program 2003 were required to complete a participant evaluation form after their involvement in the program ended. This was done in order to glean some insight on: 1) the program's overall usefulness 2) the usefulness of SREF systems to complement existing deterministic model output in the forecasting of severe

weather 3) identifying which SREF products would be useful for SPC forecasters 4) determining if the SREF output fields are of any use to an operational severe weather forecaster 5) the effectiveness of the daily operations to test the usefulness of the generation of initial perturbations approach to SREF systems 6) the comparison of the forecasts from the NCEP SREF, which utilizes traditional techniques to generate initial perturbations, with the MM5 SREF system that uses forecaster input to generate initial perturbations 7) the ability of high resolution deterministic models (Eta12, EtaKF, NMM) to predict convective initiation, structure, and mode 8) the comparison of output from the 12 km WRF model with convective parameterization and the 3 km WRF model with explicit precipitation physics as it relates to the prediction of convective initiation, evolution, and mode 9) fostering better collaboration between research and operations. The evaluation form was comprised of nine different goals, most of which were stated above, designed to evaluate the program's effectiveness in discovering new products and ideas that will assist in improving the prediction of severe weather. Each objective had goal-specific questions for the participant to rate on a scale of one to ten, with ten being the best. The ratings for each question were then averaged and a weekly mean was calculated for each question. In addition, a mean rating for each goal was computed. Comments were also given after each question. These were collected and assigned three different numbers: one for the week the participant worked, another was a job number, and the final number was the rating the participant gave the question. For job number, a one was assigned if the participant was an operational forecaster. While two, three, and four were assigned to administrators, research scientists, and others, respectively. The three numbers were assigned to investigate any correlation between how a question was

rated, when the participant wrote the comment, and what the participant's job description is.

b) Results

Reviewing the results from the participant evaluation forms illustrated that the majority of the goals of the Spring Program 2003 were accomplished well. Collaboration between operations and research were vital to the success of the program. The participants' averaged response to the program's ability to facilitate collaboration is an 8.43 out of 10, with most responses in the "very well" category. More spread was encountered when the question about the usefulness of SREF systems to complement existing deterministic model output in the forecasting of severe weather. A 6.74 (moderately effective) rating out of 10 was given for that question. Most participants in the program remarked that the program had considerable impact upon their better understanding of SREF concepts and/or utility in forecasting severe convection. According to the evaluations, the use of the perturbation approach to SREF systems was not very effectively tested by program. For this reason, not a lot was done to examine the differences in the uses between the NCEP approach and the perturbation approach. One of the goals of the Spring Program was to foster better communication between research and operations. By the end of each week, most visitors expressed the program did an excellent job of this. Most of the other goals on the participant evaluation deal with the SREF system or the deterministic models and those questions are answered in their appropriate portion of this paper.

5.) Conclusion

The 2003 Spring Program was able to accomplish the objectives that were set out before the program began. My results show that the SREF output did have a small, positive effect on the Day 2 outlook. It takes time to understand how to utilize new products, so a small, yet positive result is encouraging. I also found, through statistical analysis, that just because a model's QPF is initially misplaced or missing, does not mean the model should be discounted as a tool in the prediction of severe weather in later periods, as it pertains to watch lead time. Three of the five deterministic models examined experienced a higher mean raw score and mean rank in the second period (2100-0000 UTC) than in the first period (1800-2100 UTC). The evaluation of the participant surveys showed that a majority of participants felt the program facilitated the collaboration of researchers and forecasters well and most would be willing to work on another project like the Spring Program in the future.

ACKNOWLEDGEMENTS

The author would like to thank Steven Weiss and David Bright of the SPC for, not only the opportunity to work on such an important and timely topic, but also for the assistance and guidance provided for this study. It has been a tremendous experience. The author would also like to thank the REU project director, Daphne Zaras, for her time, effort, assistance, and overall enthusiasm. Your help this summer will never be forgotten. To the rest of the REU/ORISE students, thank you for making this summer an extremely awesome event. To Jack Kain, of the NSSL, thank you for helping me fine-tune my project. To Harold Brooks of the NSSL, thank you for the advice. I would also

like to thank all the SPC forecasters who allowed me to look over their shoulders. This work was funded by NOAA's National Weather Service through the Oak Ridge Institute for Science and Education (ORISE). My sincere thanks to the NWS EEO Office, the National Centers for Environmental Prediction's Storm Prediction Center, ORISE, and University of Oklahoma for the opportunity to participate in such a valuable experience.

References

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 379–398.

Kain, J. S., M. E. Balwin, P. R. Janish, S. J. Weiss, M. P. Kay, and G. W. Carbin, 2002: Subjective Verification of Numerical Models as a Component of a Broader Interaction between Research and Operations. *Weather and Forecasting.*

Fig. 1 – Mean Raw Score for All Periods

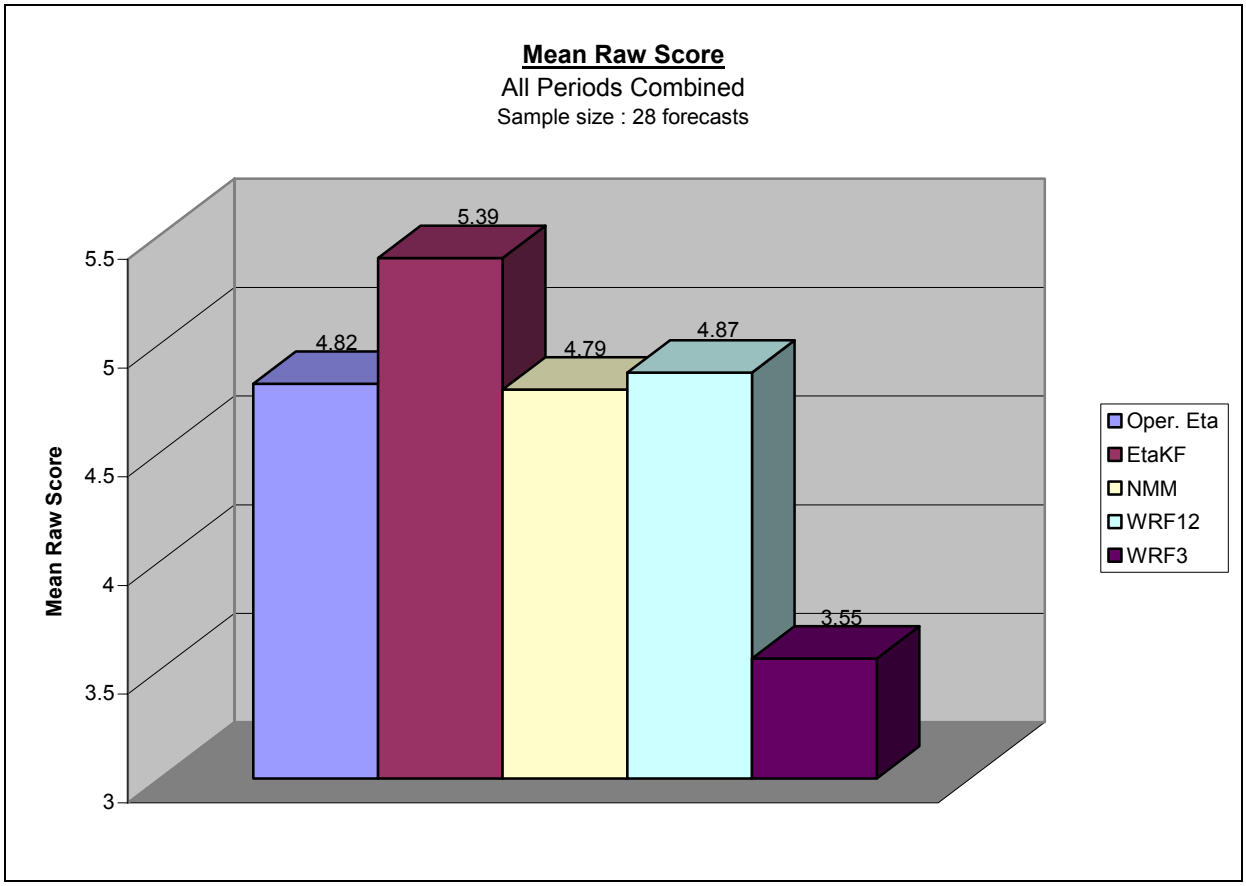

Fig. 2 – Mean Rank for All Periods

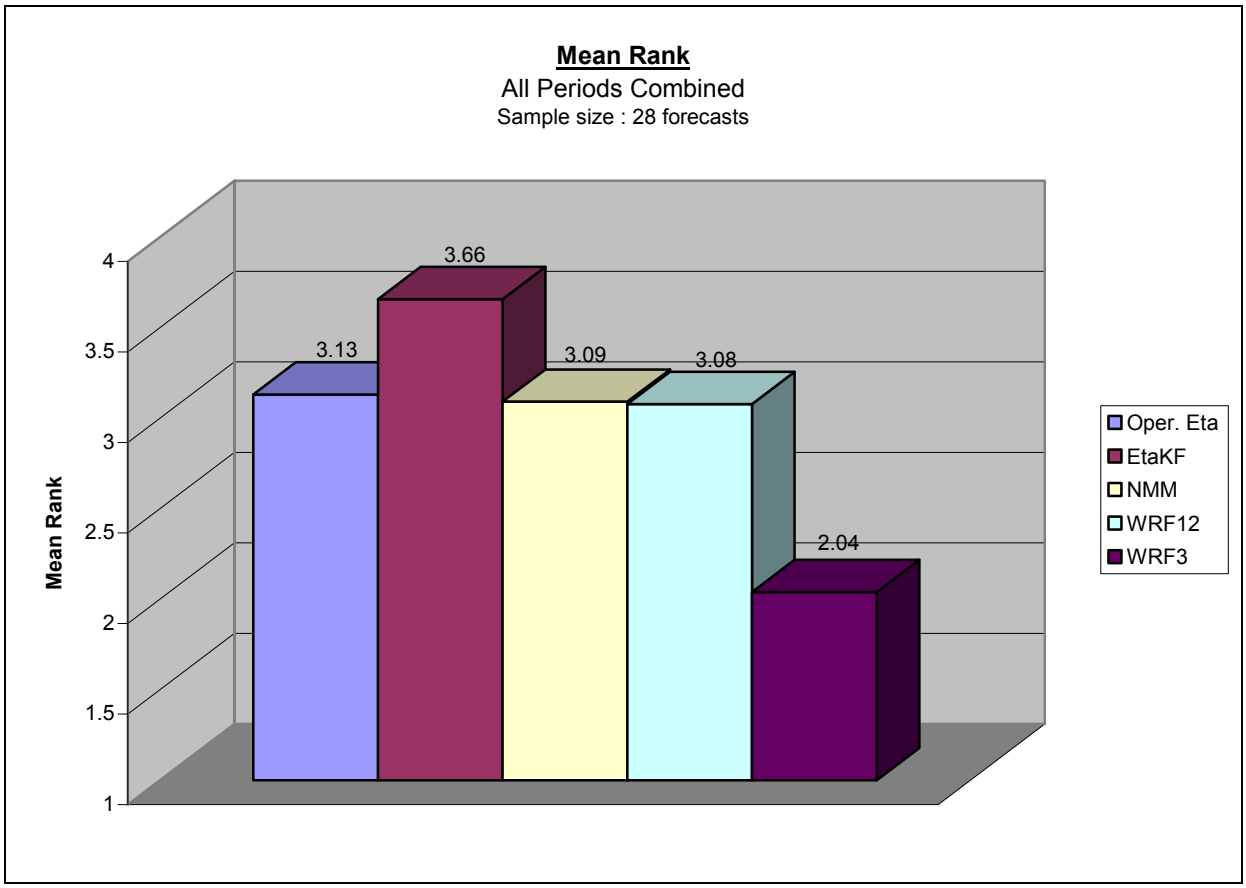

Fig. 3 – Mean Raw Score – 1st and 2nd Period Comparison

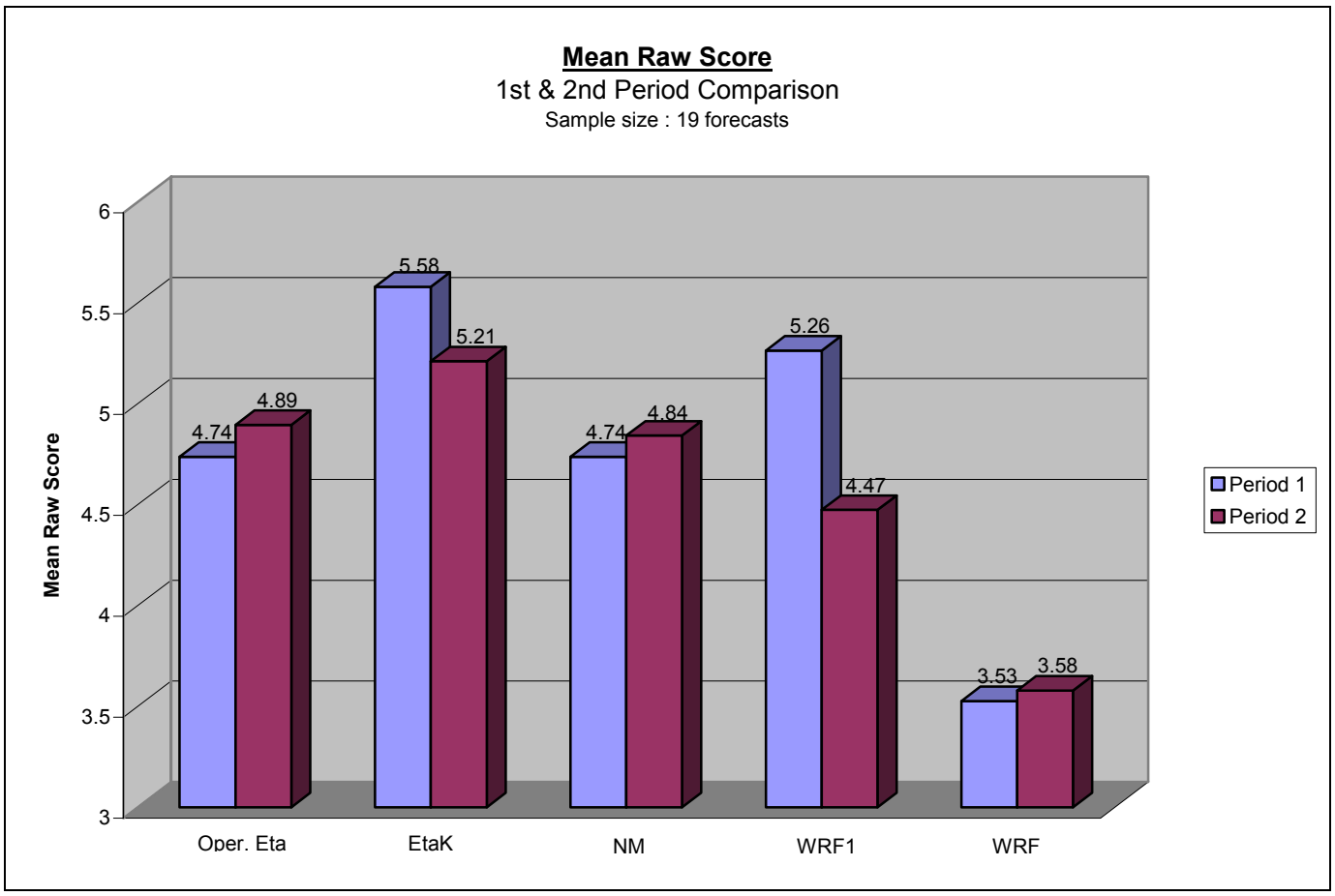

Fig. 4 – Mean Rank – 1st and 2nd Period Comparison

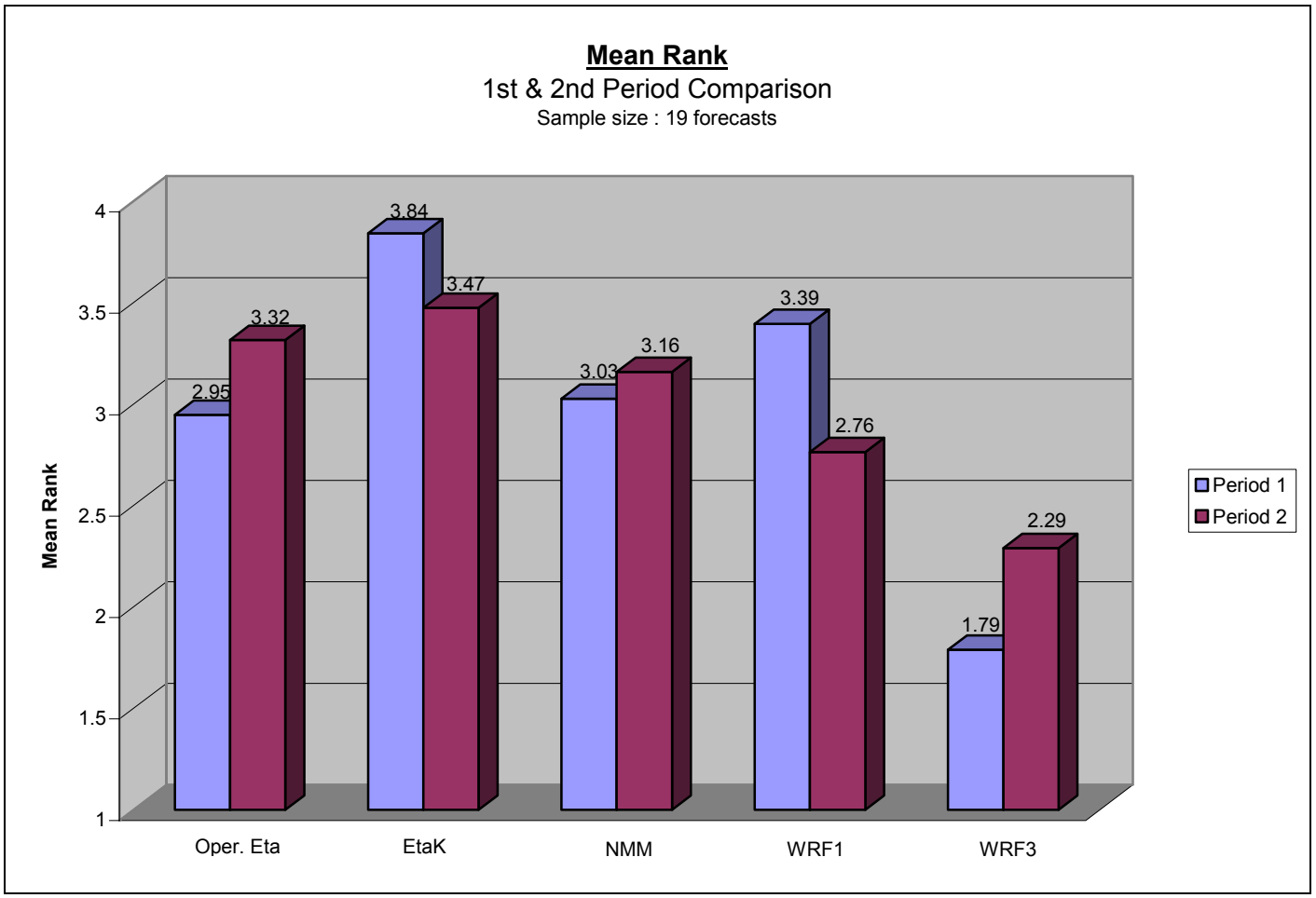

Fig. 5 – Scatterplot of Eta12 – NMM 1st Period

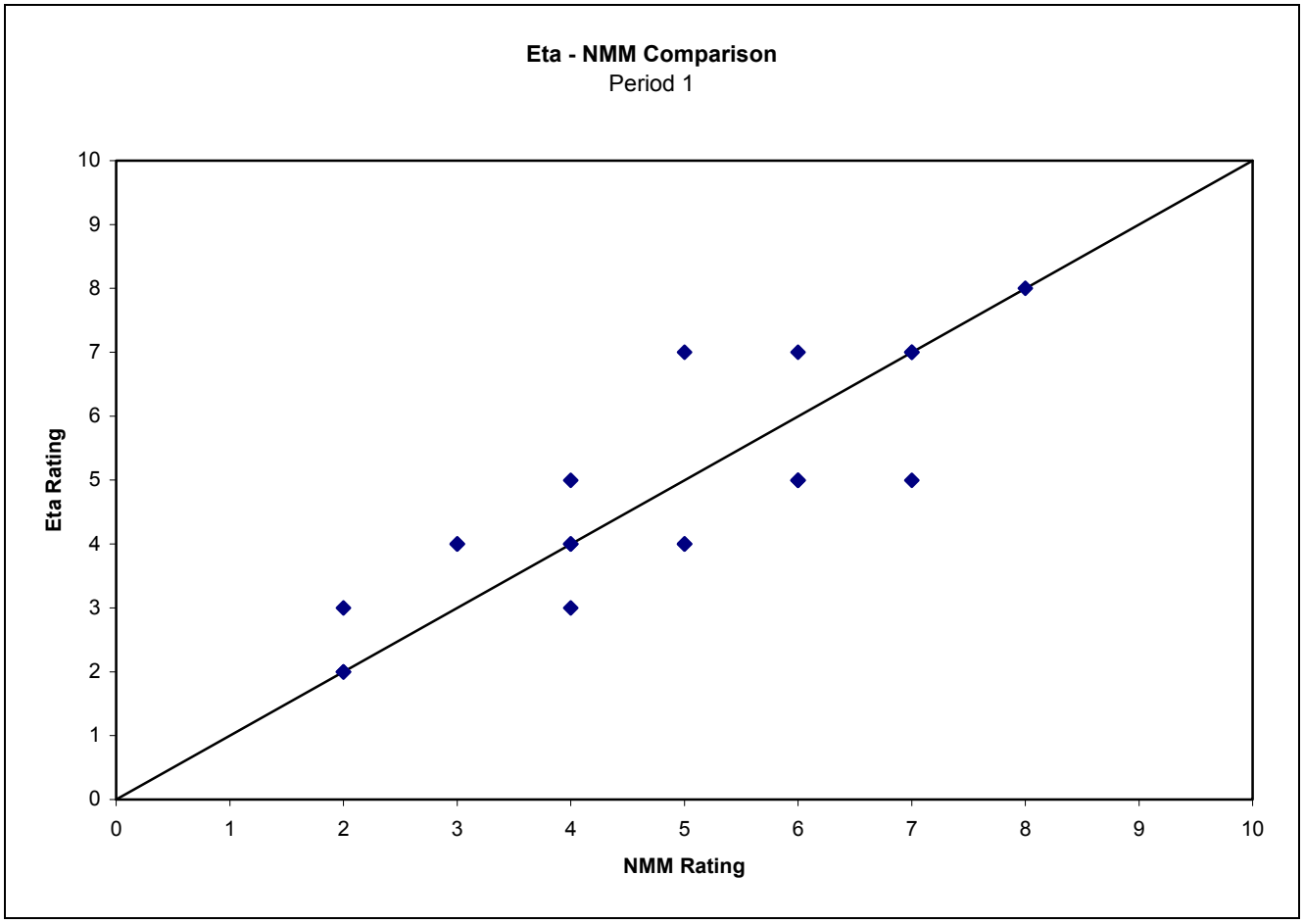

Fig. 6 – Scatterplot of Eta12 – NMM 2nd Period

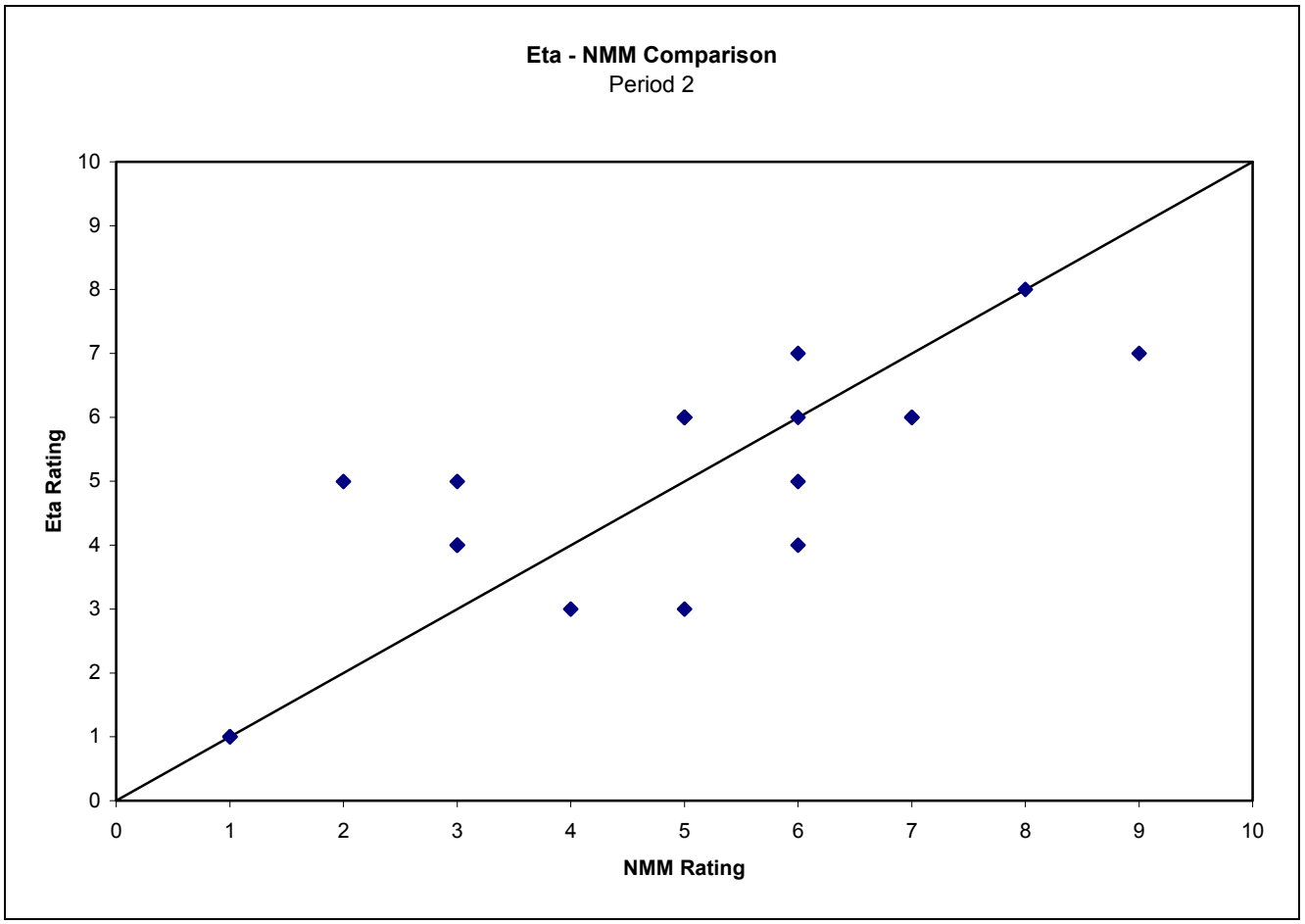

Fig. 7 – Scatterplot of EtaKF – WRF12 1st Period

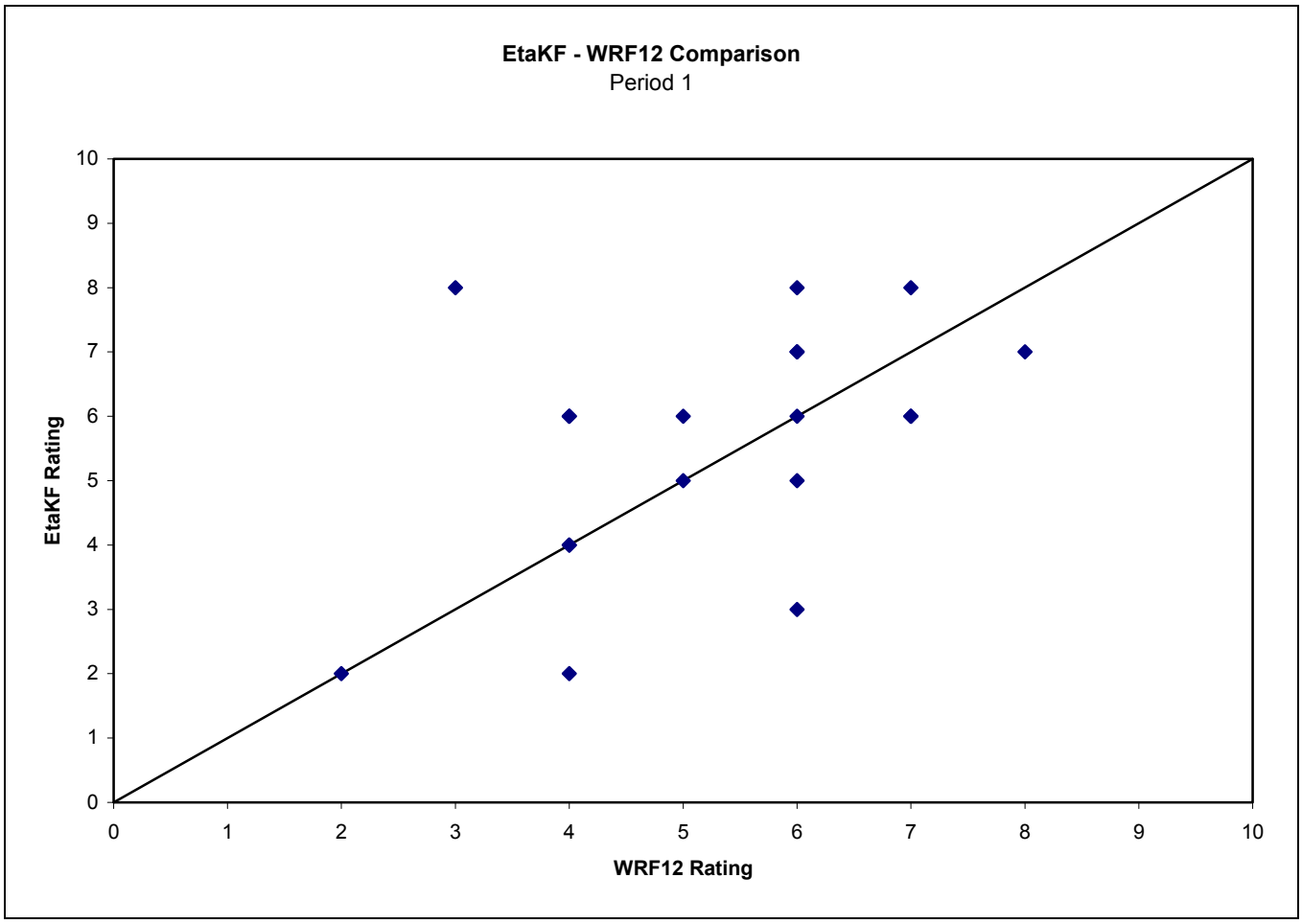

Fig. 8 – Scatterplot of EtaKF – WRF12 2nd Period

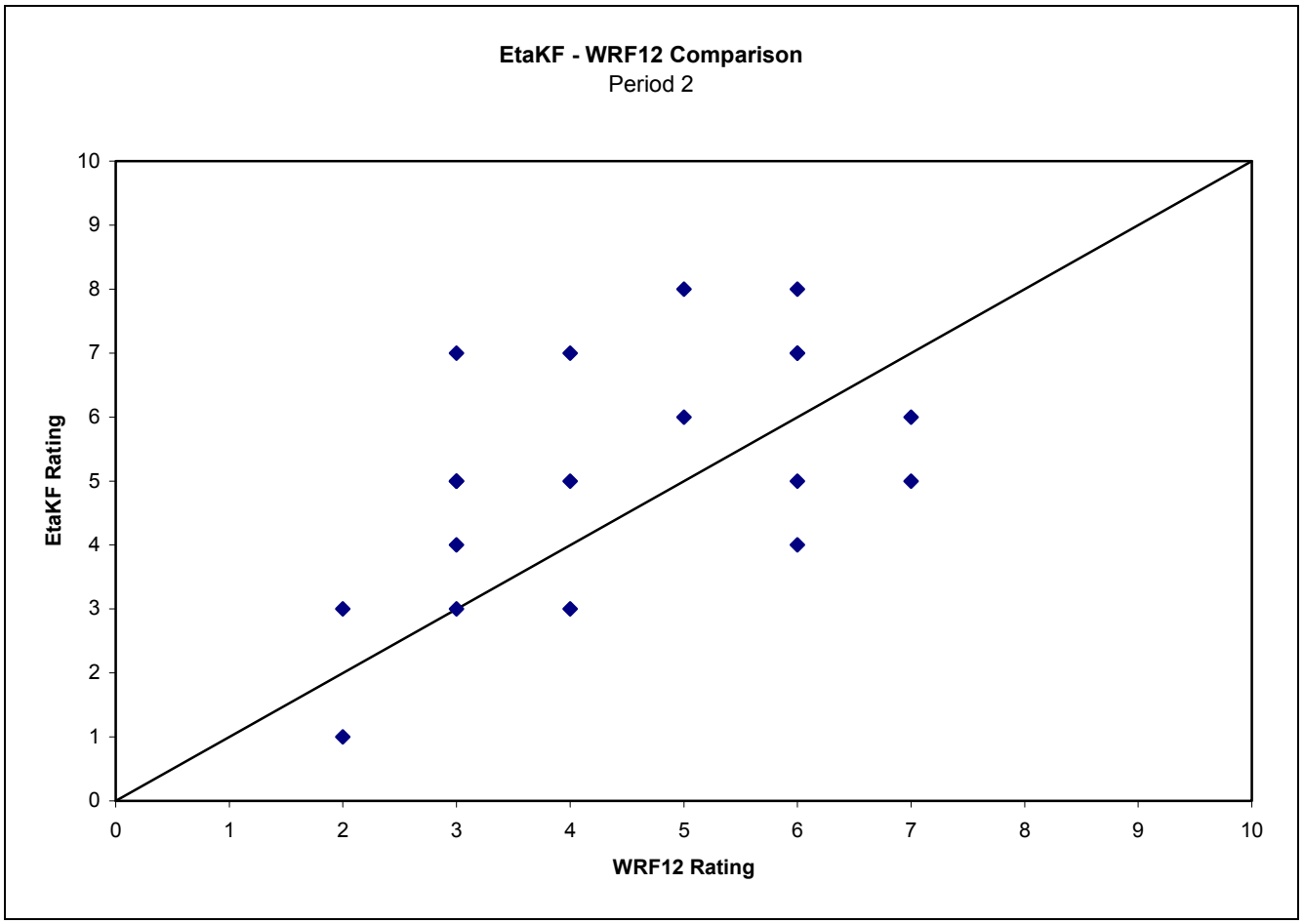

Fig. 9 – Scatterplot of WRF3 – WRF12 1st Period

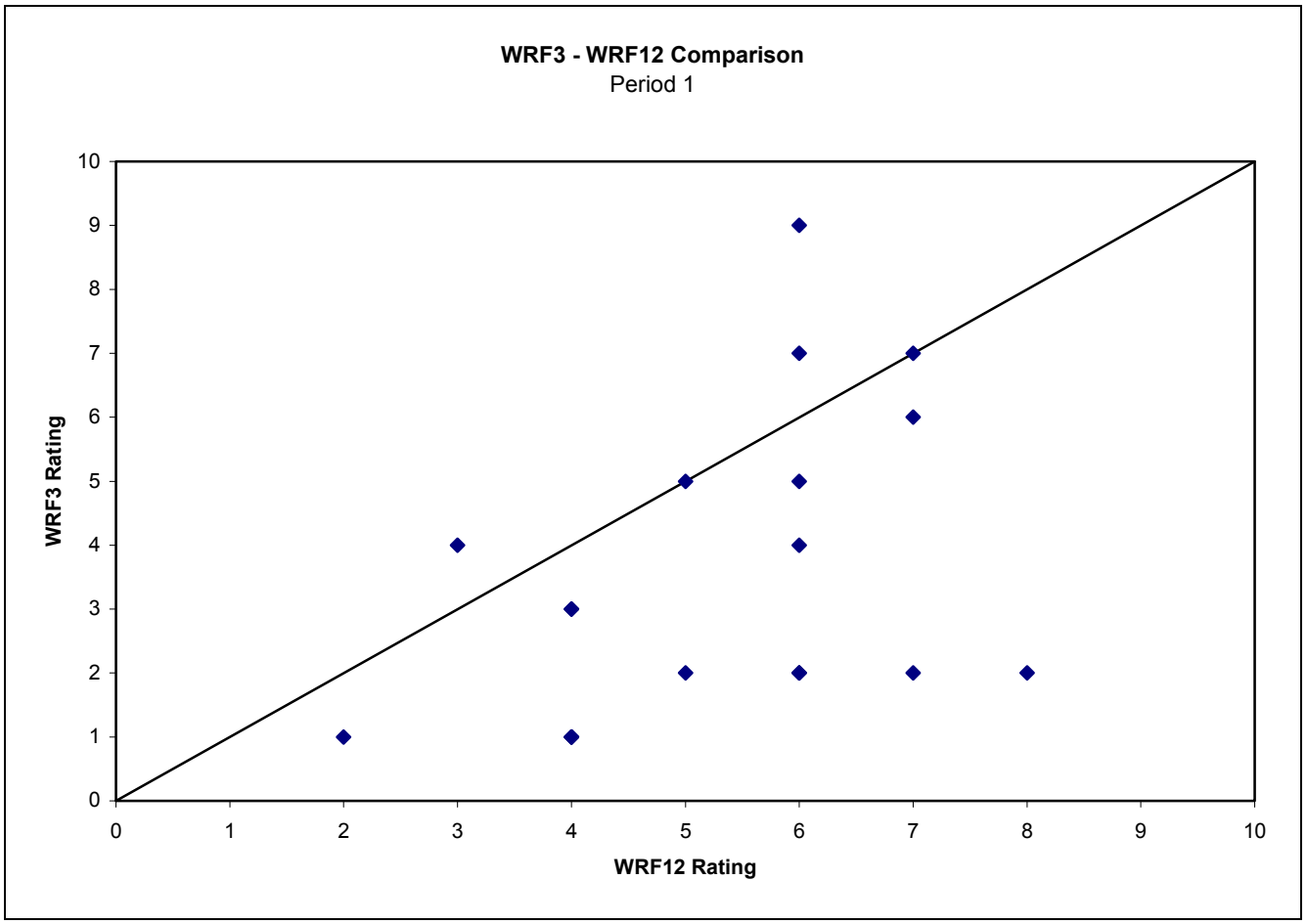

Fig. 10 – Scatterplot of WRF3 – WRF12 2nd Period

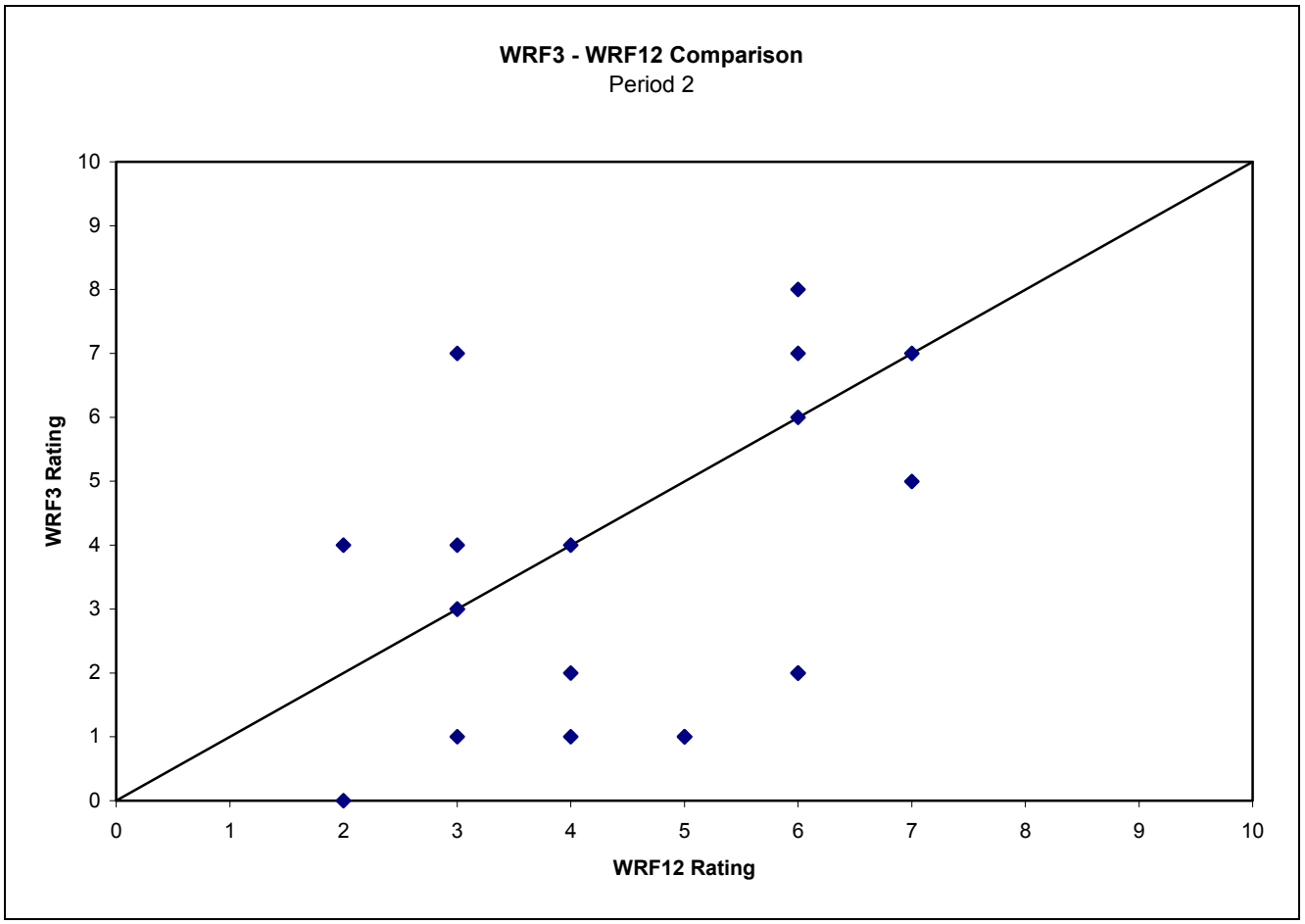

Fig. 11 – Scatterplot of Eta – EtaKF 1st Period

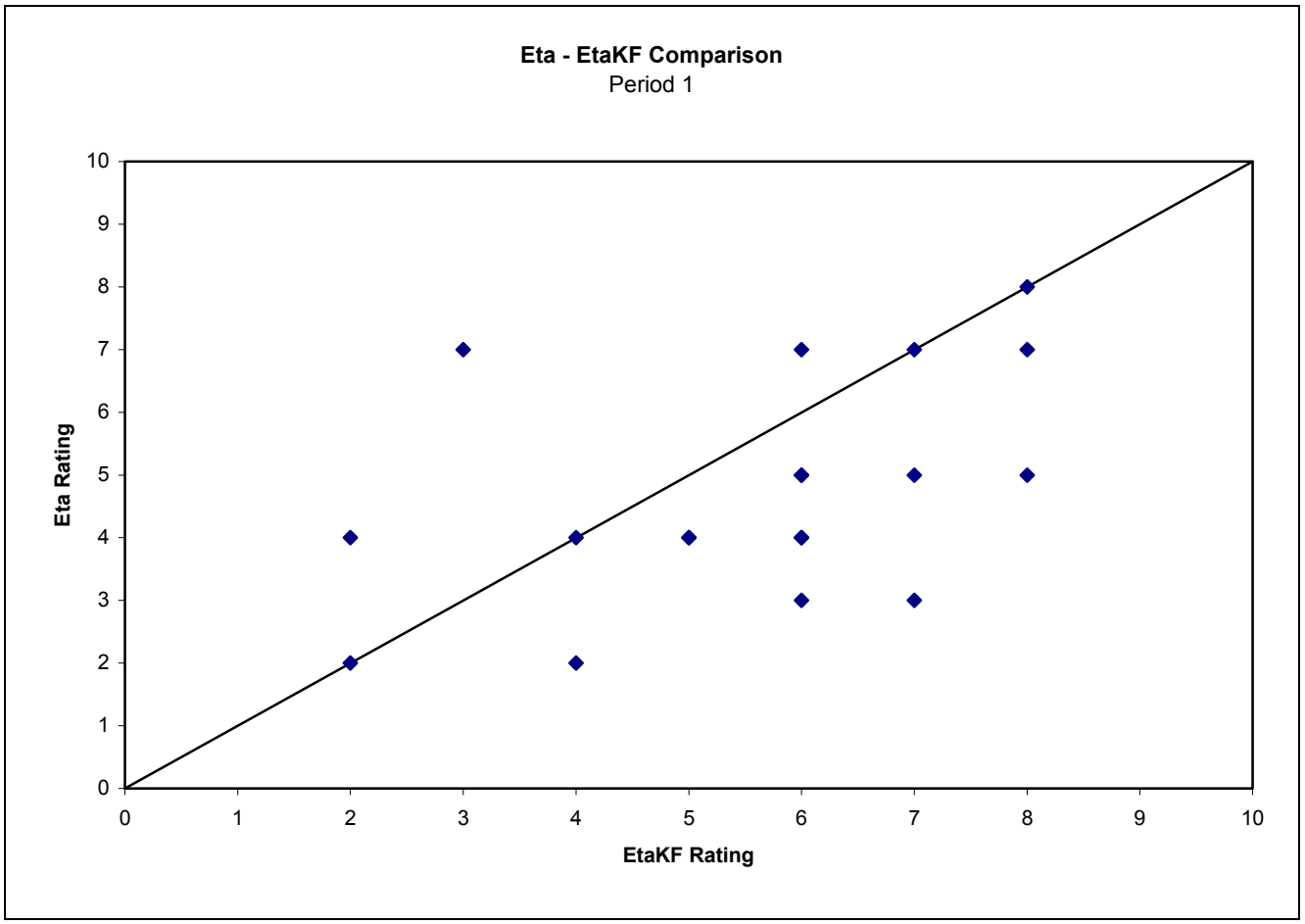

Fig. 12 – Scatterplot of Eta – EtaKF 2nd Period

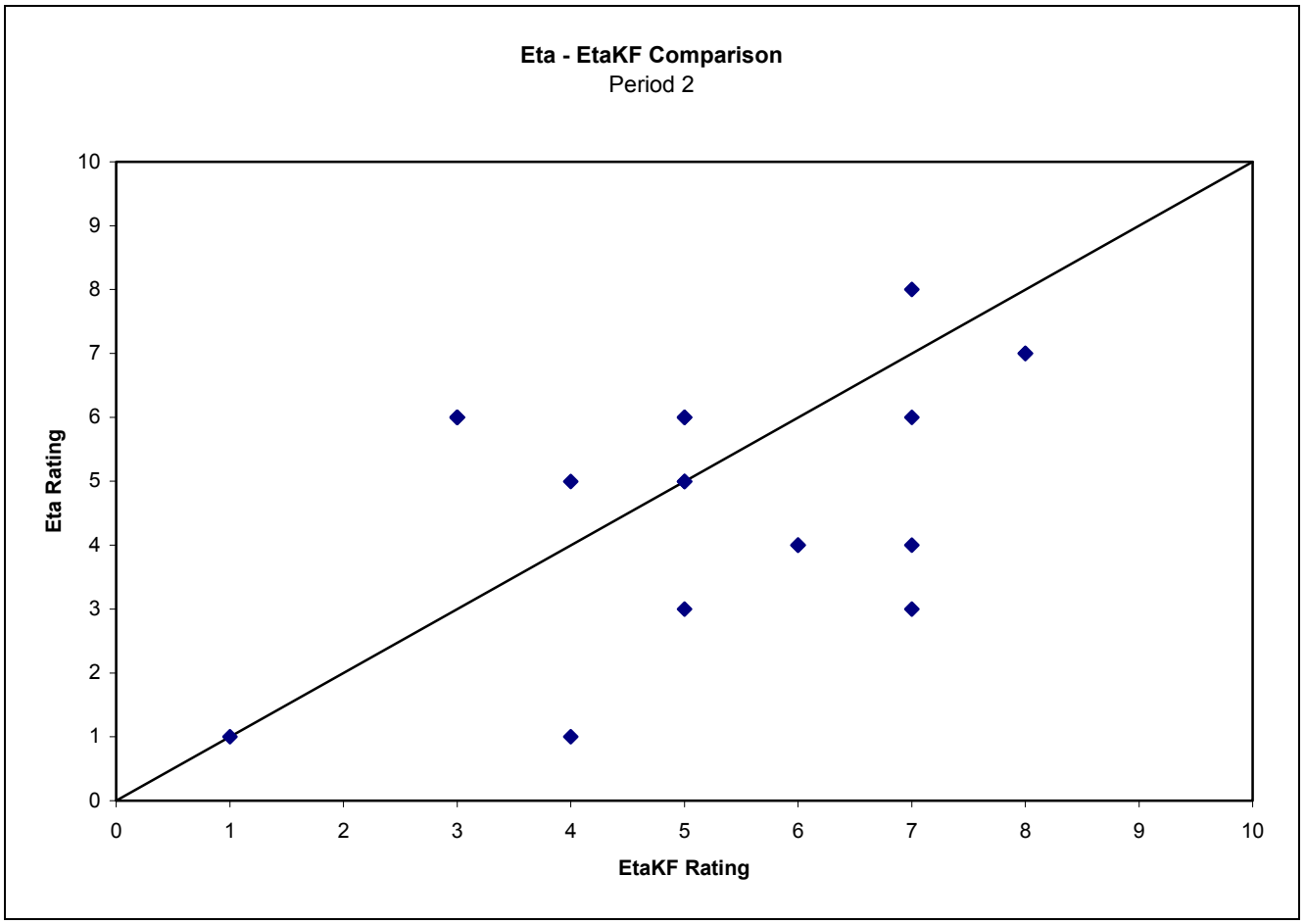

Fig. 13 – Scatterplot of Eta Comparison - Both Periods

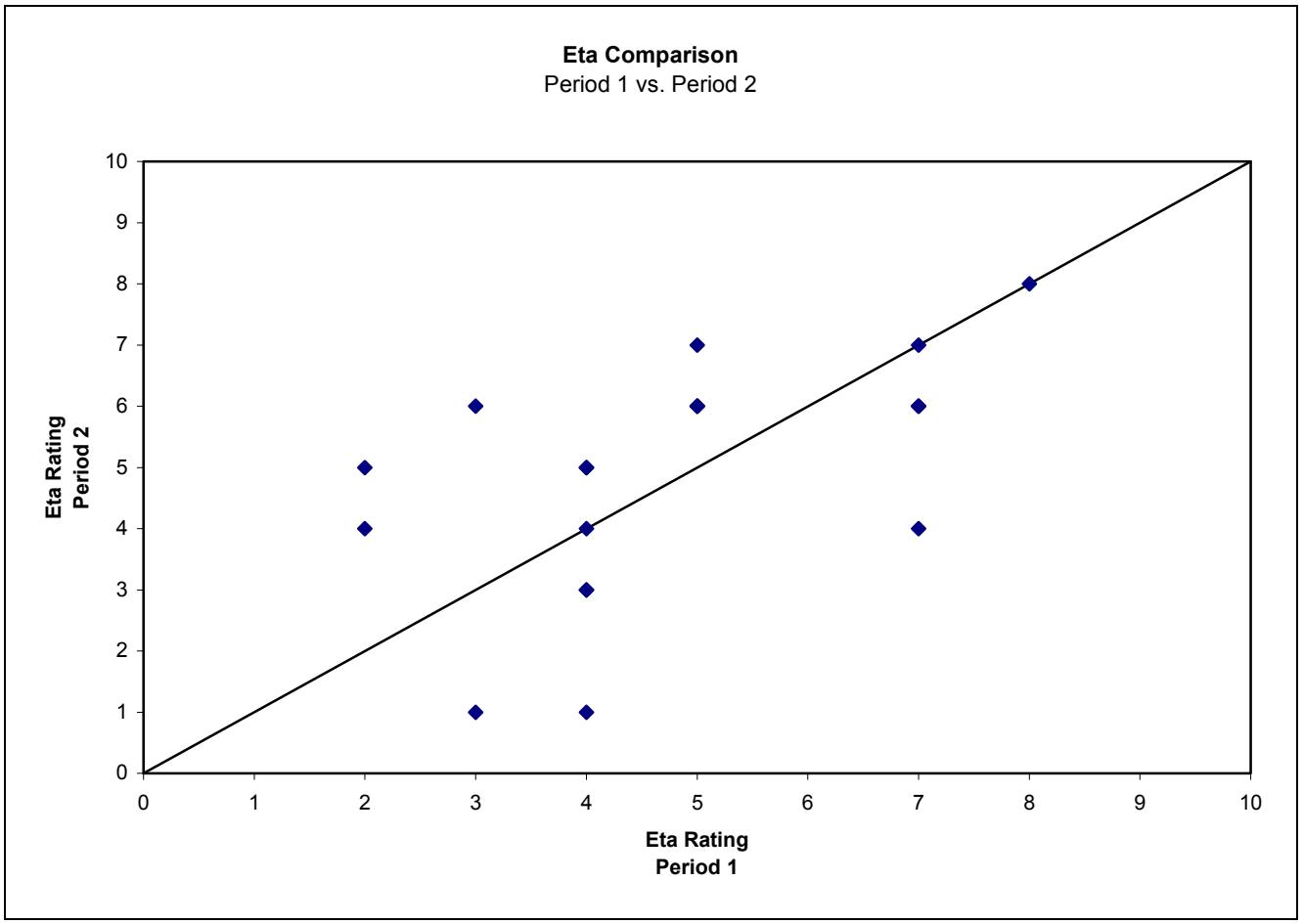

Fig. 14 – Scatterplot of EtaKF Comparison - Both Periods

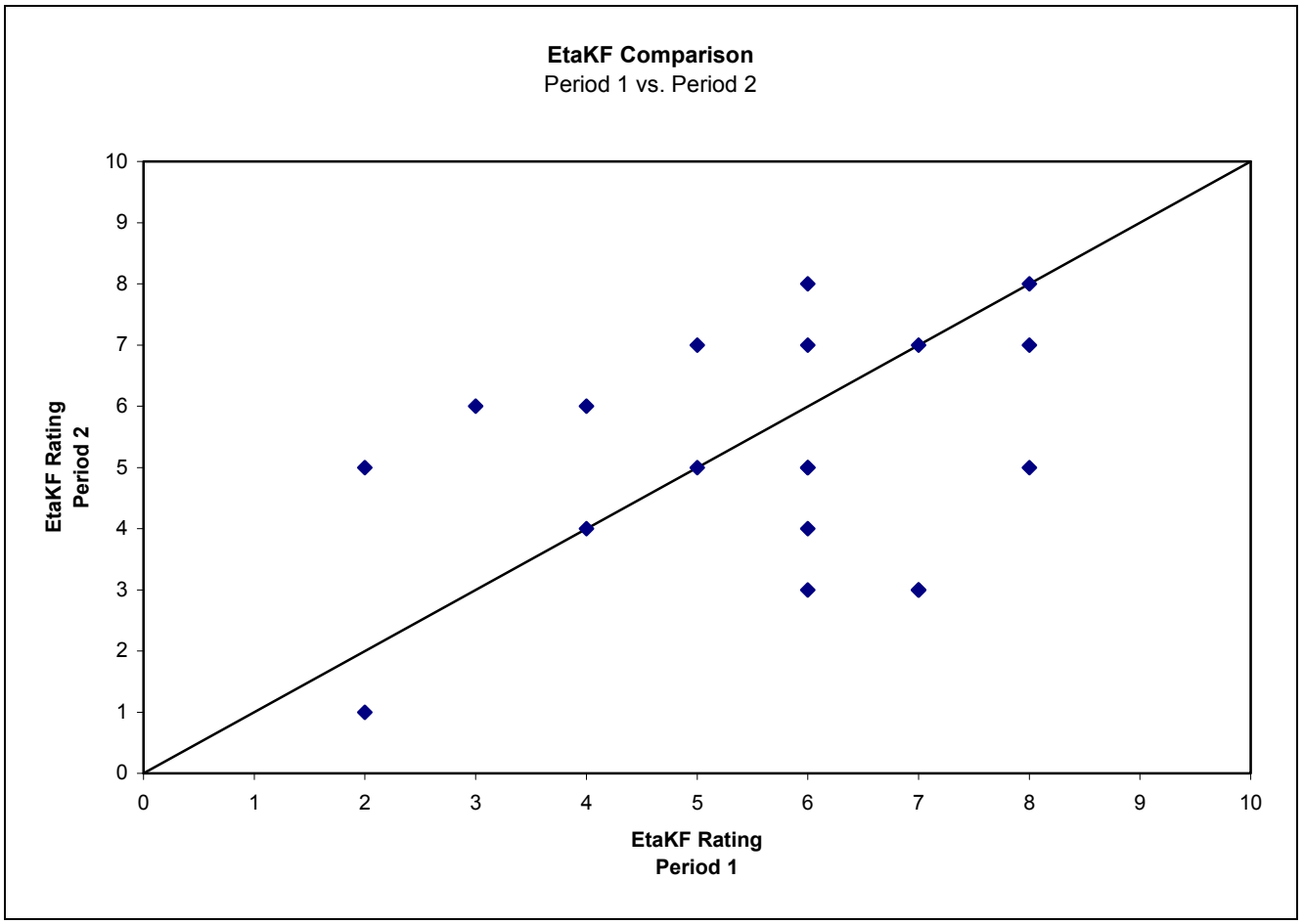

Fig. 15 – Scatterplot of NMM Comparison - Both Periods

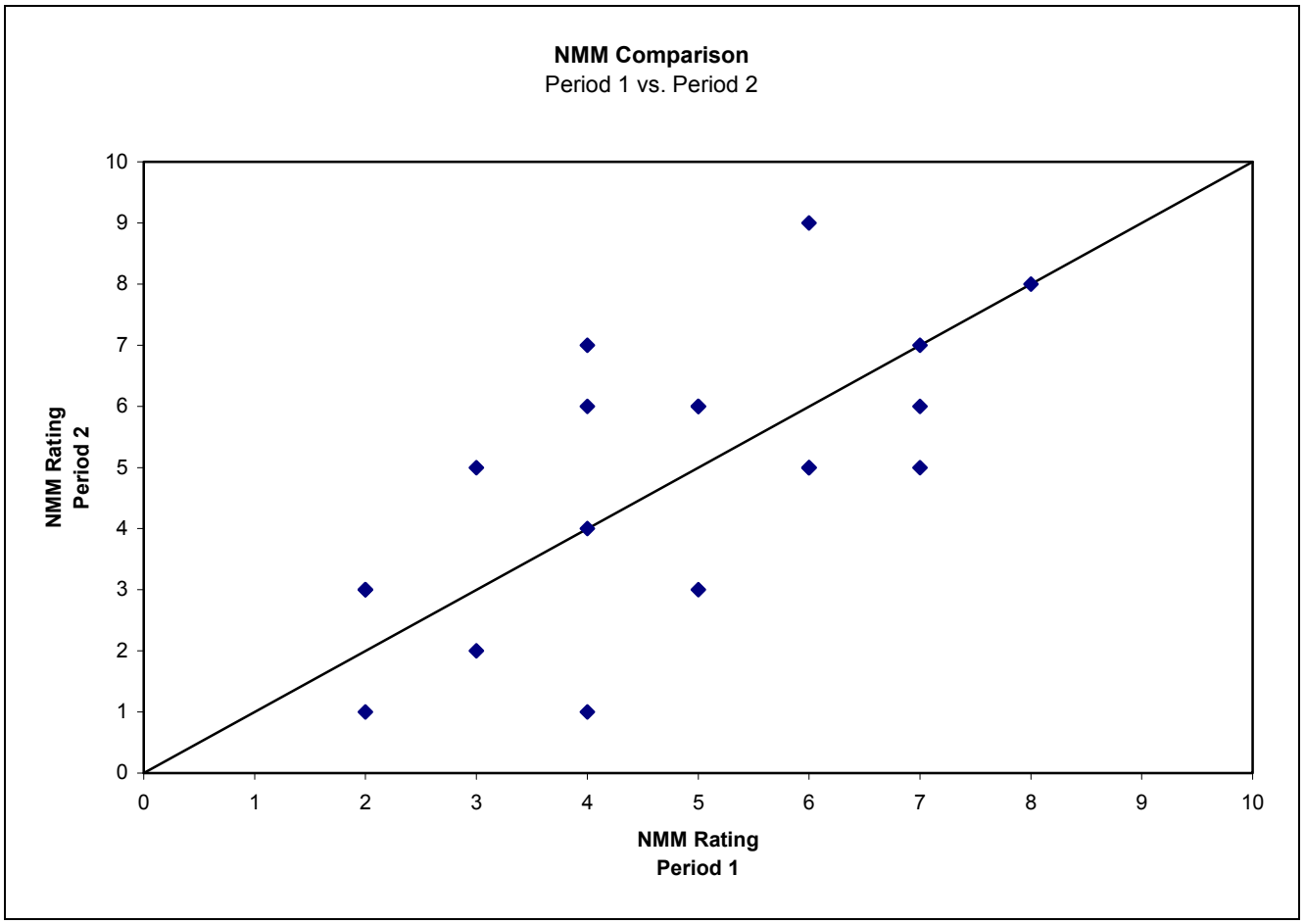

Fig. 16 – Scatterplot of WRF12 Comparison - Both Periods

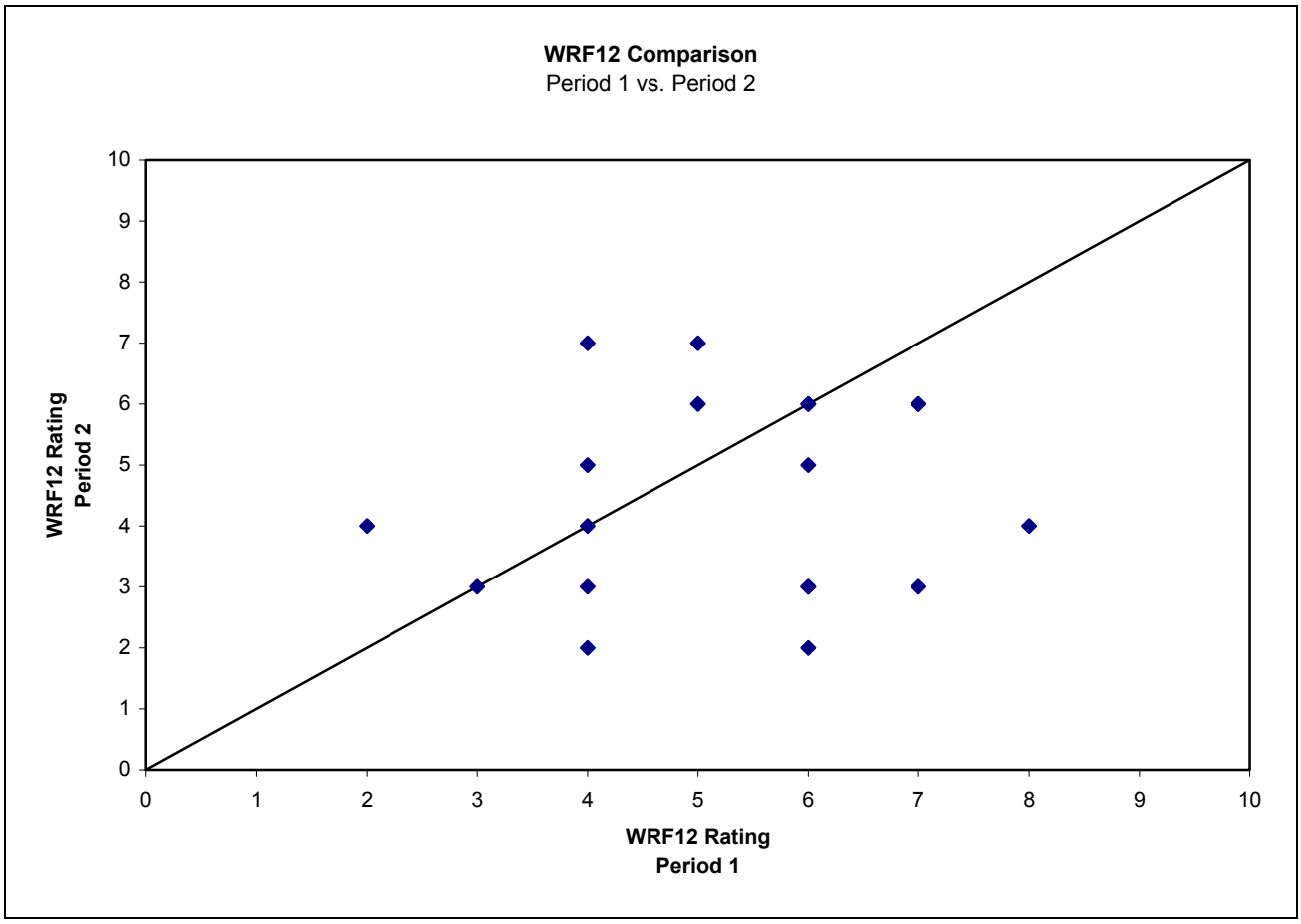

Fig. 17 – Scatterplot of WRF3 Comparison - Both Periods

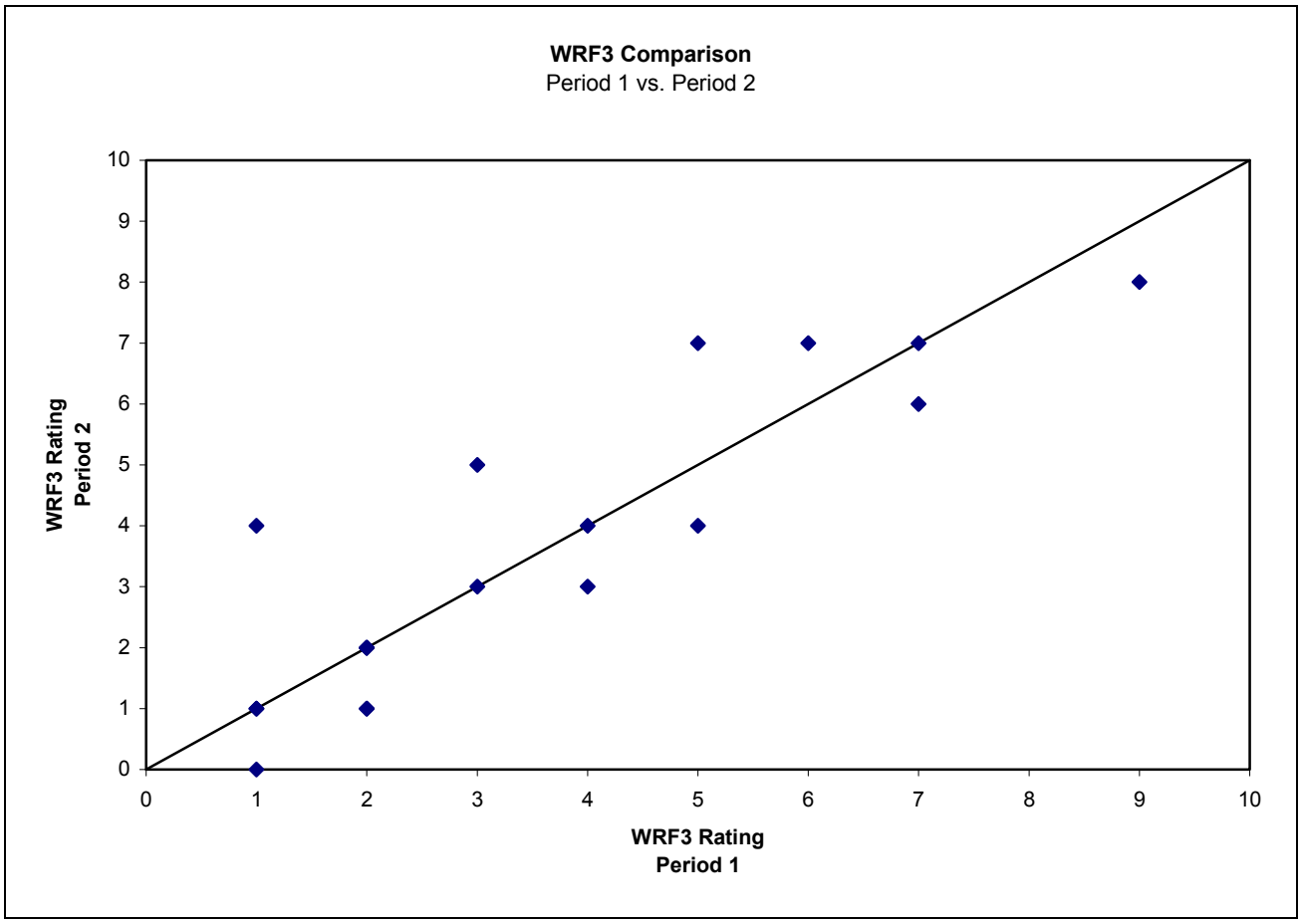

Fig. 18 – T-Test Scores

T-Test - All Periods					
	Oper. Eta	**EtaKF**	**NMM**	**WRF12**	**WRF3**
Oper. Eta	X				
EtaKF	0.078	X			
NMM	0.893	0.069	X		
WRF12	0.882	0.070	0.850	X	
WRF3	0.005	0.000	0.016	0.001	X

T-Test - Period 1					
	Oper. Eta	**EtaKF**	**NMM**	**WRF12**	**WRF3**
Oper. Eta	X				
EtaKF	0.064	X			
NMM	1.000	0.042	X		
WRF12	0.220	0.446	0.287	X	
WRF3	0.050	0.001	0.043	0.004	X

T-Test - Period 2					
	Oper. Eta	**EtaKF**	**NMM**	**WRF12**	**WRF3**
Oper. Eta	X				
EtaKF	0.517	X			
NMM	0.871	0.492	X		
WRF12	0.461	0.079	0.593	X	
WRF3	0.055	0.033	0.141	0.115	X

T-Test - Period 1 vs. Period 2						
		Period 1				
		Oper. Eta	**EtaKF**	**NMM**	**WRF12**	**WRF3**
Period 2	**Oper. Eta**	0.702	0.091	0.674	0.497	0.041
	EtaKF	0.370	0.473	0.428	0.927	0.026
	NMM	0.818	0.069	0.790	0.448	0.098
	WRF12	0.656	0.047	0.690	0.122	0.117
	WRF3	0.104	0.003	0.102	0.011	0.848