Long Range Sea Ice Drift Model Verification

Robert W. Grumbine *

NOAA/NWS/NCEP/EMC/MMAB Contribution no. 3NN

* *Corresponding author address:* Robert W. Grumbine, #2108 5830 University Research Ct., College Park MD 20740

E-mail: Robert.Grumbine@noaa.gov

ABSTRACT

Integrated sea ice drift is predictable to about 16 days in the Arctic. This surprising result is an extension from Grumbine (1998), where there was no apparent decline of skill through the 6 days forecast lead at that time. Forecasts from 1998 to 2007 provide a further test of that, and a continued search for good measures of model skill. Index of Agreement, used in Grumbine (1998), turns out to be a poor skill measure for sea ice drift.

1. Introduction

The drift of sea ice is important for safety of navigation and fishing. A feature of particular importance is the motion of the ice edge. Since most vessels are not strengthened for working inside the ice pack, even low concentrations (15%) of thin (0.1 m) ice are a concern [Cavalieri et al. (1991)]. Alaska Region National Weather Service forecast offices, lead by Anchorage, have been making use of sea ice drift guidance models for a number of years (at least since 1978 [Crisci (1978)]) to assist in making these forecasts. The nature of these models has been to use 'virtual floes'. That is, if there were an ice floe at this location at the start of the model run, then it is predicted by the model to drift this distance in the following direction. The lead time of these models was 2 days in 1978 [Crisci (1978)], to 4 days some time before 1989, and to 6 days in 1989 [Grumbine, inspection of source code] where it remained until January 2001.

In 1997 a new drift model was implemented operationally [Grumbine (1998)], hereafter G1, which again went to day 6. The new model was superior to the old, measured by drift error radius. It was a sufficient improvement that Anchorage Weather Service Forecast Office soon thereafter requested that guidance be extended [Page, pers. comm, 1998] to day 10. Since the atmospheric model used by the drift model, the Global Forecast System (GFS), extends to day 16, the experiment was begun running the model to day 16. The experiment was successful, and the 16 day guidance became operational in January 2001.

A remarkable feature of the verification statistics for forecasts of up to 6 day lead in G1 was that there was no apparent decline in skill by any of the measures used – index of agreement, correlation, vector correlation. As noted in G1, this behavior could be expected in a case where the steadily increasing bias, from biased forcing, was offset by a decreasing random error component as positive errors are offset by negative errors. One of the interests for this paper is to determine how long that compensation could continue. We will see that that limit is 2-5 days, depending on skill measure.

The new model has been running as a parallel since 14 April 1993, and operationally since September 1997, which gives us over 15 years worth of experience.

We have two main questions: Does the model have skill beyond day 6? and Is there potential for a higher resolution model to improve on this one? The first can be tested in a fairly straightforward manner. The second, we will examine by comparing the verification statistics between buoys as a function of initial distance between buoy and forecast point. If the model performs better for buoys closer to forecast points, then we have reason to believe that we should go to a higher resolution model. This resolution test applies only to the quantity tested – N-day integrated drift. This point is important to keep in mind as there is reason to believe that more than averaged drift conditions are required in order to model the sea ice thickness correctly [e.g. Geiger (1997)].

2. Skill Assessment Methods and Results

We will again use the International Arctic Buoy Program (IABP) observations of ice motion for our verification, partly to ensure comparability with the G1 scores.

As in G1, we use several different assessors of skill. We use several because each will penalize different types of errors. Also, we are seeking skill measures which provide insight to how well the model is performing. We expand here from four assessors to ten. In considering how well the model performs by each measure, we are simultaneously also examining how informative each measure is. We will wind up concluding that only five are needed, and one of the four (index of agreement) used in G1 is not useful for evaluating this model.

Since the GFS Hybrid [Lord et al. (2007)] implementation on 1 May 2007 the drift model has actually been running with a regression tuned with only 2 weeks of model 10 meter winds, rather than geostrophic winds of G1 and prior models. For consistency as we examine performance measures and model skill, only the 1998-2007 frame is used in this paper. A subsequent paper will examine the 2007-present model. And will use 2 years of observations for tuning.

Distance Correlation

One measure is simply the linear correlation between the forecast distance of drift and the observed distance. This is a statistic most are familiar with, and with well-known weaknesses. One is that it will consider a 50 km drift forecast correct even if the drift is southward rather than northward. The other is that the magnitude of drift error is considered equally significant regardless of how large the observed drift is. That is, if the forecast is 22 km, and the observed drift is 2 km, this will be penalized the same (in a least squares linear regression) as a forecast of 40 km versus an observation of 20 km. Both are 20 km wrong, but the relative error is very different.

The distance correlation between forecast and observed for each forecast lead (averaging period in this model) is shown in figure 1. The peak is at 3 days, with it not declining below the 1 day lead's score until day 8.

$$slope = (\Sigma xy - n * \overline{x} * \overline{y})/(\Sigma xx - n * \overline{x} * \overline{x}) \tag{1}$$

$$intercept = \overline{y} - b * \overline{x} \tag{2}$$

$$correlation = (\Sigma xy - n * \overline{x} * \overline{y})/\sqrt{}(\Sigma xx - n * \overline{x} * \overline{x})/\sqrt{}(\Sigma yy - n * \overline{y} * \overline{y}) \tag{3}$$

$$\tag{4}$$

Slope of Regression

The slope of the regression line between forecast and observed distances provides information beyond the correlation itself. If the model were perfect, this slope would be 1, as would correlation. But one can have a perfect correlation with a model that is continually wrong by a factor of 2. This parameter will show that situation. As we see in figure 1, the model consistently under-predicts the drift of ice, at most about 65% of the observed. Given this model's linear nature, an *a posteriori* fix could be made to the model's output. As we will see in discussion of the intercept (below), part of this is systematic bias in the model's guidance.

This skill measure peaks for 2-3 days, and declines below the 1 day forecast, slightly, at day 4. Of the figure 1 skill measures, it shows the shortest period to peak skill, and for sustaining day 1 skill in to the future.

Regression Intercept

In making the linear regression for the slope, above, we also find the optimal intercept for the line. This is the bias in forecast drift distances. Regression intercept is displayed in figure 2, where we see it increase to about 45 km at 16 day forecast lead.

Vector Correlation

Vector correlation [Crosby et al. (1993)] will penalize directional errors as well, although it is still subject to the usual problems of correlation scores. Figure 1 shows this score as well. A perfect score is 2 (2 dimensional vectors). This score peaks at 4 days lead/averaging, and does not decline below the 1 day lead's skill until 11 days. This vector correlation's definition for a sample is:

$$r^2 = Tr[\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}] \tag{5}$$

where Tr denotes the trace of the matrix and Σ_{ij} is:

$$s(u_i, u_j) \quad s(u_i, v_j) \tag{6}$$

$$s(v_i, u_j) \quad s(v_i, v_j) \tag{7}$$

s() is the sample covariance, and i, j = 1 represent the observations and i, j = 2 are the forecasts.

Index of Agreement

Index of Agreement [Wilmott et al. (1985)] ignores direction, as does correlation, but does include a sense of how large the error is relative to the forecast. Figure 1 shows this as well, this skill having little trend, whether to rise or fall in time. Given the behavior of all other measures, this is inconsistent, and the score shows little ability to distinguish better from worse. As such, it should not be used as a skill measure for integrated ice drift distances.

$$d_2 = 1 - [\Sigma\omega_j|\mathbf{e_j}|^2]/[\Sigma\omega_j(|\mathbf{p_j} - \overline{\mathbf{o}}| + |\mathbf{o_j} - \overline{\mathbf{o}}|)^2 \quad (8)$$

where d_2 is the index of agreement, summations are from 1 to N (the number of observations), ω are weights to correct $\mathbf{e_j}$ for being over- or underrepresentative, $\mathbf{e_j}$ is the error in the jth forecast, p is the prediction, o is the observation, and $\overline{\mathbf{o}}$ is the mean of the observations weighted by ω. In our case, the weights are taken to be unity. The index of agreement will be largest when the numerator is smallest (forecasts agree with observations), and when the denominator is largest (large natural variability - the o_j vary greatly from $\overline{o}$).

RMS Distance Error

The root mean square of the distance error prevents under-forecast drifts from compensating for over-forecast drifts. In the limit of an unbiased model, this becomes the standard deviation of the errors. As for the regression intercept, figure 2 shows this monotonically increasing in time. It is always greater than the intercept, increasing to about 52 km at 16 day forecast lead.

Error Radius

The error radius is the difference between the position of the drifter (ice floe) at the end of N days versus the forecast location. This is a very natural figure in terms of visualization.

It is also shown in figure 2. The mean shows monotonic increase, to about 35 km at 16 days. The root mean square of this error also shows monotonic (aside from a curious decline at 16 day lead, likely artefactual) increase to about 60 km at 16 day lead.

The preceding 4 measures are all distances, all of which increase monotonically with time. One can equate them, then, to a speed, which ranges from 2-5 cm/s.

RMS Direction Error

Figure 4 shows the root mean square of the direction error. We see it decline to about 71 degrees at 5 day forecast lead. It remains superior to the day 1 forecast through 9 days. Errors increase from the minimum at day 5, though only slightly.

We can compare this measure to the result of a random guess for direction, given in equation 1. Uniformly random selection of direction would have an rms error of 103 degrees. At 16 days, the directional error is still less than this. Actually, as shown in figure 4, the rms direction error at day 16 is only slightly larger (85 degrees versus 71 degrees) than at day 5.

$$\sqrt{\frac{1}{180}\int_0^{180} \theta^2 d\theta} = 103^\circ \tag{9}$$

Mean Distance Error

Figure 3 shows the mean distance error, allowing underforecasts to compensate for overforecasts. This shows an increase through time much like the measures above which do not permit compensation. This confirms the impression from the slope of the regression line that the model is systematically biased.

Mean Direction Error

The mean direction error, also shown in figure 4, is consistently small, untrended, and of the same sign – a magnitude of -5 to -10 degrees. While small, it is the same order of magnitude as the drift rotation in the first place (8 degrees in G1).

3. Implications for Resolution of Velocity grid

The mean error radius is given in table 1 with respect to matchup radius. That is, for the given radius, buoys which were this close to the Skiles point (the set of points used by **?** were treated as drifting in the same distance and direction as a buoy which really was at the Skiles point. Table 1 also gives the error radius for only those matchups which were in the annulus between the given matchup radius and the next smaller matchup radius. Thus, in the row for 55 km (the matchup radius used in G1), values for the matchups are given both for initial distance between forecast point and buoy being between 0 and 55 km, and between 38.9 and 55 km. Figure 5 helps illustrate this.

The mesh of Skiles points is a 381 km polar stereographic grid. This rectangular mesh does not translate uniquely to a representation based on radius from a point. There are three obvious methods of translation:

1) Tile the plane with circles inscribed inside the rectangles (leading to radius = dx / 2, where we let dx be the 381 km spacing). This leaves gaps between the circles, especially along the diagonals between grid points.

2) Tile the plane with circles having the same area as the squares, (giving radius = dx / $\sqrt{\pi}$). This gives both some gaps along the diagonals, and some overlap perpendicular to them.

3) Tile the plane with circles circumscribed around the squares. This gives radius = dx / $\sqrt{(2)}$, and no gaps, but much more overlap.

As compromise among the three, I will take the circles having the same area as the squares. In this case, the matchup radius of 55 km, for instance, corresponds to a grid spacing of just about 100 km (97.5), much finer than the model's actual grid spacing of 381 km. Table 1 also lists this equivalent grid spacing.

In order to test for statistically significant differences in the difference between the means of two populations having unknown variance, the test statistic is [e.g. p 283 Devore (1982)]

$$t = \frac{(mean(x_1) - mean(x_2))}{\sqrt{(\frac{s_1^2}{m} + \frac{s_2^2}{n})}} \tag{10}$$

where n = number of observations from population 1, m is the number from population 2, x is the variable of interest, subscripts referring to which population was sampled, and s^2 is the sample variance. This is also given in table 1, for annulus versus succeeding annulus. By comparing annuli, we have independent samples. There these are t statistics, with (n+m-2) degrees of freedom. A two tail t test for p = 0.95 has critical value of 1.97 for N tending to infinity.

So, considering first the annulus from 165 to 190 km, representing the outer area of the model's grid, we find no (statistically) significant improvement until the matchup radius 55 km (annulus 38.9-55 km). The next improvement is with the matchup radius of 27.5 km (annulus 19.5 to 27.5 km). By this point, the annuli have far fewer observations, down to about 500 from the outermost annulus' 13,000.

In terms of grid spacing, this suggests that there is no particular reason – in terms of modeling N day integrated sea ice drift – for a model's grid spacing to be reduced from 381 km until it can be reduced to about 100 km. And no need for spacing to be reduced from 100 km until it can be reduced to about 50 km. Given modern computing this is of mostly historical interest for stand-alone sea ice models, but is relevant for coupled air-sea-ice models, where, for instance, the most recent Climate Forecast System [Saha et al. (2011)] includes a half degree, about 50 km grid spacing, ocean and ice model. The point is worth examining with a denser model and observational data set, and for other model types where it may be prohibitive to perform full testing at the higher resolution.

4. Conclusions

Our first question, whether the model shows skill beyond 6 days' lead is answered yes. It is clearly at least better than guessing random directions for ice floe drift. The second question, whether the skill can be improved by increasing model resolution, is also answered yes. Approximately a 25 km grid spacing is supported by this analysis, and such a model was submitted for NCEP operational implementation 21 May 2012. This model will also produce kml [Consortium (2012)] output for geographic inspection.

Skill does indeed, eventually, decline. Depending on the measure used, peak skill is some time from day 2 to 5, with 4 days being a compromise lead between the different measures. Also, insofar as forecasters consider the model to be useful/skilled at 1 day lead, it shows at least equal skill for forecast leads of 4-10 days. How long, exactly, depends on the measure used.

We also see that Index of Agreement, mean direction error, and mean distance error are not useful skill measures for sea ice drift as they vary so little that we cannot distinguish between day 1 and day 16 forecasts, even though the model, according to the other measures, clearly does vary in skill with respect for forecast lead. The magnitude of the mean direction error suggests, however, that the model could be improved markedly by correct tuning of the angle difference between geostrophic winds and ice drift.

Five of the measures essentially repeat each other, so do not shed additional light on the model's behavior. These are the mean distance error, the rms distance error, the error radius, and the regression intercept. Of these, error radius has some advantages for physical interpretation.

There are four additional measures which provide unique information with respect to each other and the error radius regarding sea ice drift model. These are the vector correlation, distance correlation, the slope of the regression between forecast and observed drift distance, and the rms direction error. So I suggest that sea ice drift model verification use 5 measures – error radius, vector correlation, distance correlation, the slope of the regression between forecast and observed drift distance, and the rms direction error.

Acknowledgments.

I would like to thank Mark Ortmeyer for providing observations from the International Arctic Buoy Programme, and Xingren Wu and Arun Chawla for their helpful reviews.

REFERENCES

Cavalieri, D. J., J. P. Crawford, M. R. Drinkwater, D. T. Eppler, L. D. Farmer, R. R. Jentz, and C. C. Wackerman, 1991: Aircraft active and passive microwave validation of sea ice concentration from the defense meteorological satellite program special sensor microwave imager. *JGR*, **96**, 21,989–22,008.

Consortium, O. G., 2012: Kml standards. Open Geospatial Consortium, URL `http://www.opengeospatial.org/standards/kml`, URL `http://www.opengeospatial.org/standards/kml`.

Crisci, R. L., 1978: Objective guidance for forecasting sea ice drift. NOAA/NWS Alaska Region.

Crosby, D. S., L. C. Breaker, and W. H. Gemmill, 1993: A proposed definition for vector correlation in geophysics: Theory and application. *J. Atmos. Oceanic Technol.*, **10**, 355–367.

Devore, J. L., 1982: *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole Publishing Co.

Geiger, C. A., 1997: Importance of sea ice for validating global climate models. Tech. rep., AMS, 122-125 pp.

Grumbine, R. W., 1998: Virtual floe ice drift forecast model intercomparison. *Weather and Forecasting*, **13**, 886–890.

Lord, S., J. Derber, M. Iredell, R. Treadon, W.-S. Wu, D. Parrish, and D. Kleist, 2007: Ncep satellite data assimilation overview, 1 may 2007. NCEP/EMC, unpublished, unpublished.

Saha, S., et al., 2011: The ncep climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, **91**, 1015–1057, doi: 10.1175/2010BAMS3001.1.

Skiles, F. L., 1968: Empirical wind drift of sea ice arctic drifting stations. Tech. rep., Arctic Institute of North America, 239-252 pp.

Wilmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, D. R. L. K. M. Klink, J. O'Donnell, and C. M. Rowe, 1985: Statistics for the evaluation and comparison of models. *J. Geophys. Res.*, **90**, 8995–9005.

List of Tables

TABLE 1. Error radius skills, for 16 day lead/averaging period, for matchup radii and annuli; all distances in km

Matchup Radius	N	Mean error radius	s	dx equiv	N annulus	mean error radius annulus	t vs. previous	t vs. 190 km
19.5	488	32.2	39.7	34.6	488	32.2	-0.44	3.11
27.5	948	31.4	38.5	48.7	460	30.55	2.14	3.62
38.9	2119	34.7	46.6	69.0	1171	37.37	-0.42	1.77
55	4247	35.6	44.6	97.5	2128	36.5	1.87	2.96
77.8	8000	37.4	50.0	137.9	3753	39.44	0.17	0.99
110	15806	38.5	57.9	195.0	7806	39.63	0.47	1.05
155	29451	39.2	52.4	274.7	13645	40.01	0.68	0.68
190	42663	39.6	52.9	336.8	13212	40.49		

List of Figures

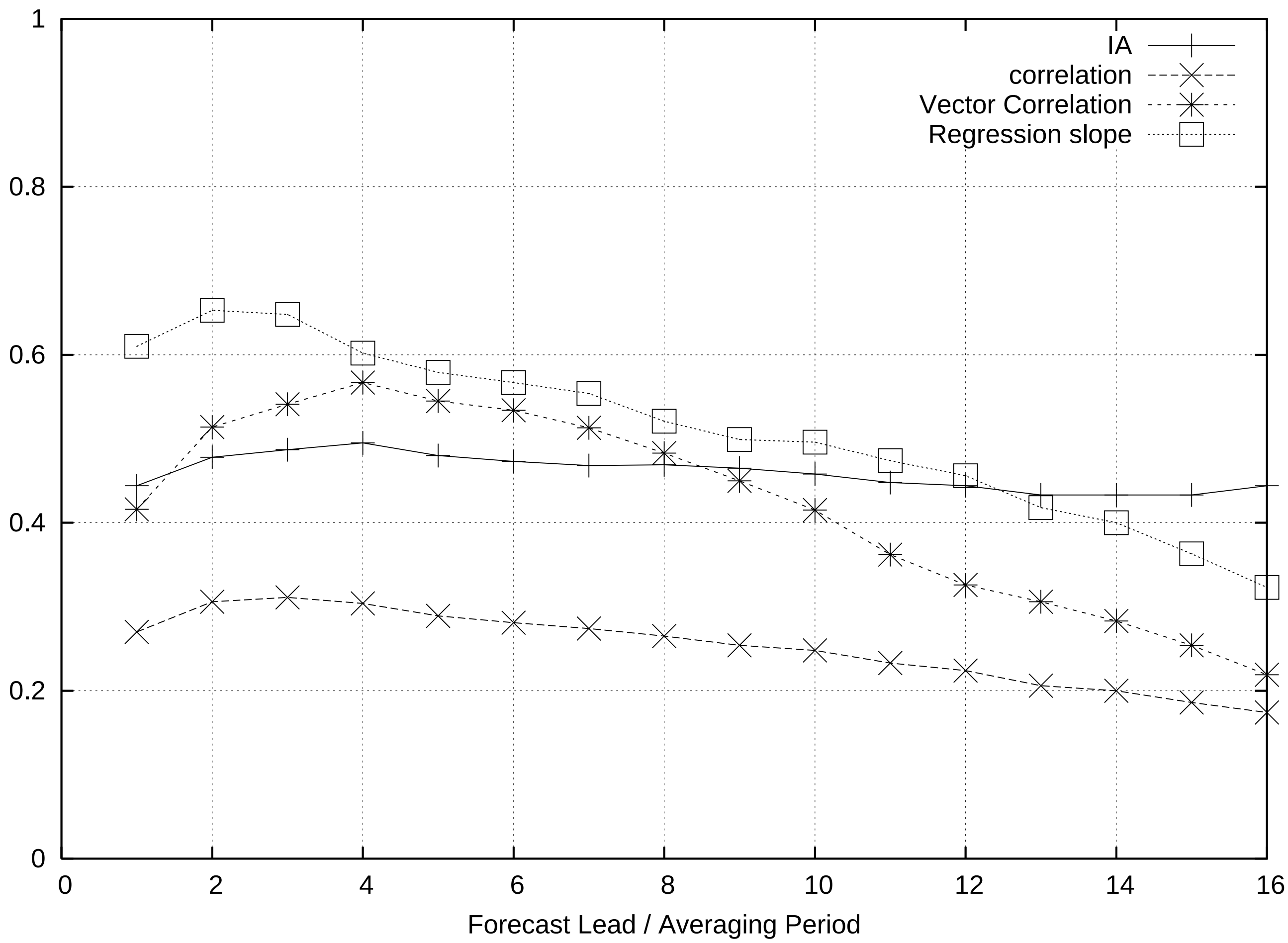

FIG. 1. Skill versus forecast lead/averaging time for index of agreement, correlation of distance, vector correlation, slope of forecast vs. observed distance drifted

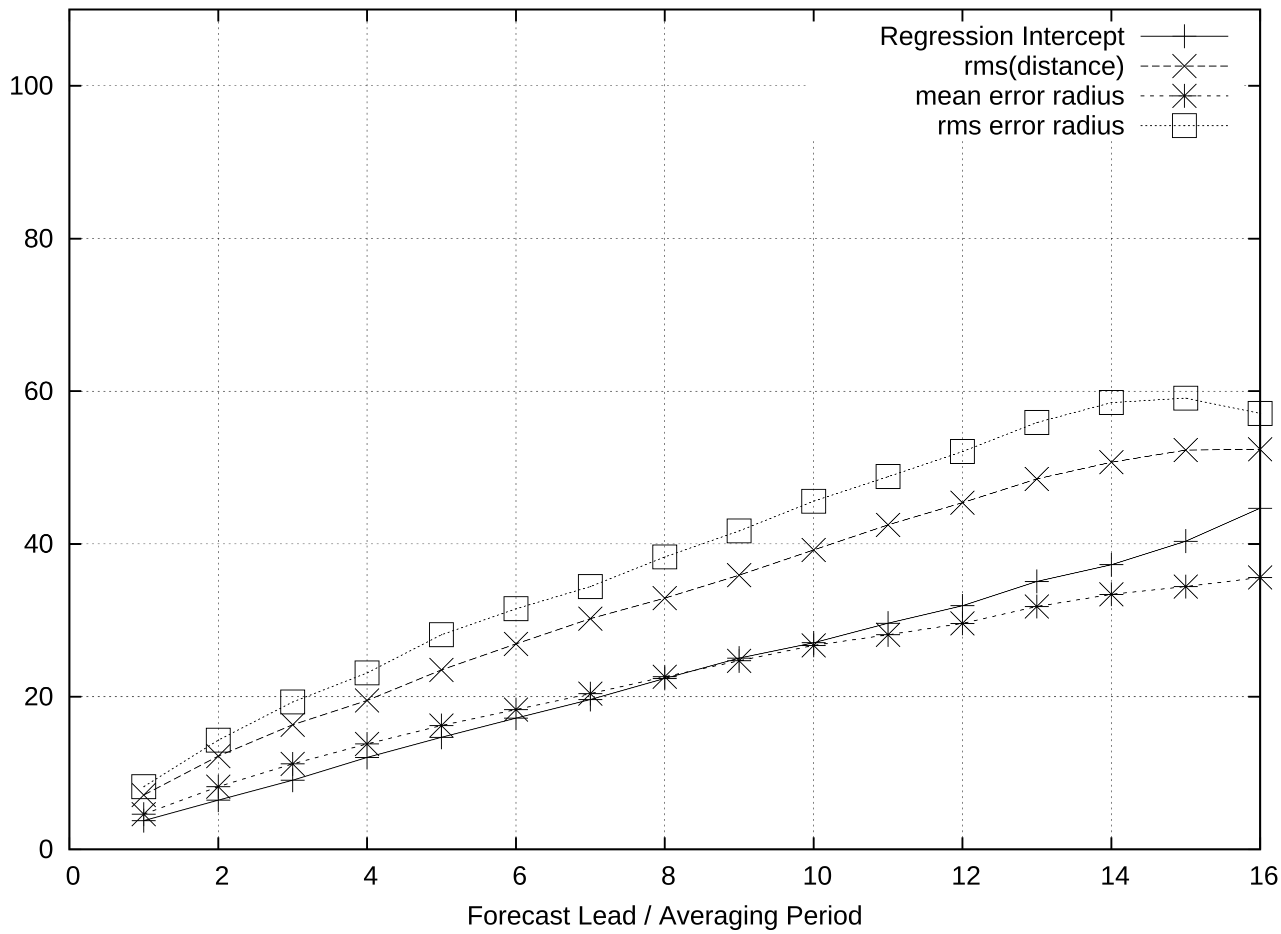

Fig. 2. Skill versus forecast lead/averaging time for regression intercept, rms drift distance error, mean error radius, rms error radius

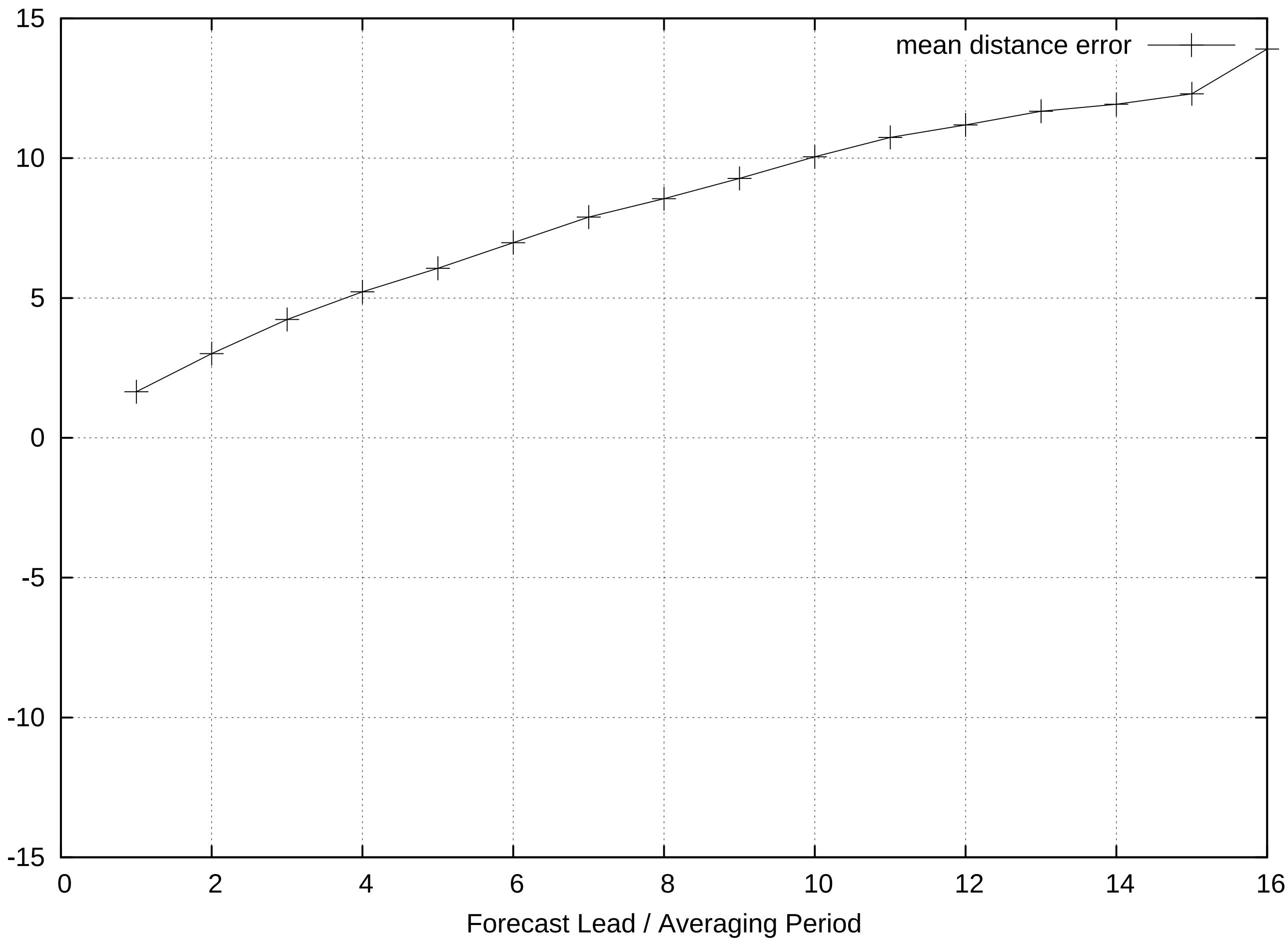

FIG. 3. Skill versus forecast lead/averaging time for mean distance error

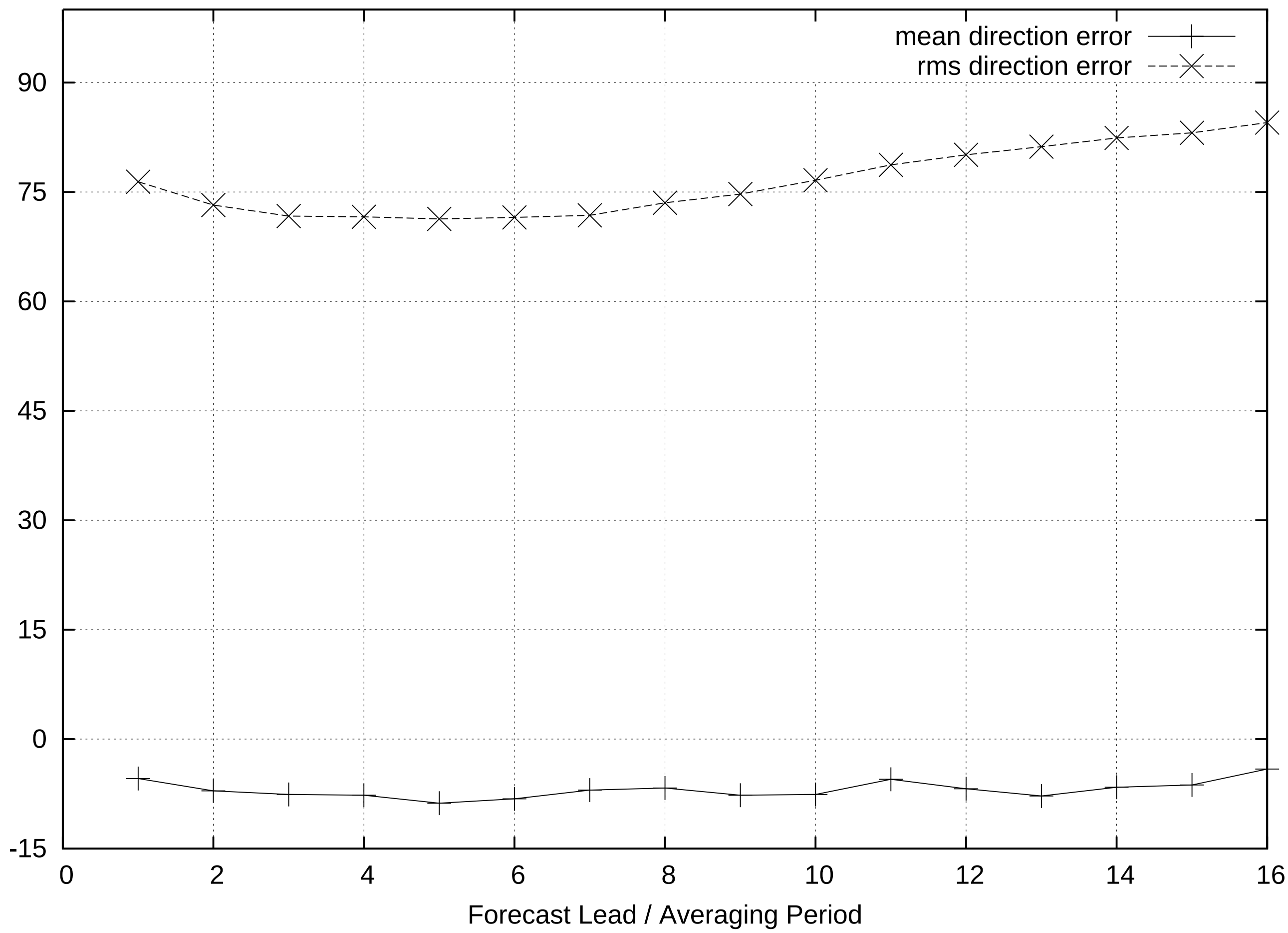

FIG. 4. Skill versus forecast lead/averaging time for mean direction error and rms direction error

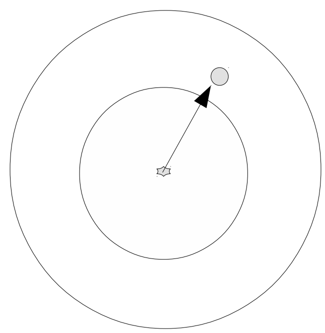

FIG. 5. Illustration of annular matchup