NMC NOTES

On the Weights for an Ensemble-Averaged 6–10-Day Forecast

H. M. VAN DEN DOOL AND L. RUKHOVETS

Climate Analysis Center, Washington, D.C.

23 March 1994 and 19 April 1994

ABSTRACT

A scheme to optimally weight the members of an ensemble of forecasts is discussed in the framework of calculating an as accurate as possible ensemble average. Results show, relative to a single member, a considerably improved 500-mb height forecast in the 6–10-day range for the Northern Hemisphere. The improvement is nontrivial and cannot be explained from simple smoothing. This method is used in operations at the National Meteorological Center.

1. Introduction

Since December 1992 the National Meteorological Center has been producing a set of five global model runs out to 12 days ahead every day. Including the multiple runs that were made yesterday and the day before (see Fig. 1), we have, every day, an ensemble of 14 members to aid in any operational forecasts in the range from 1 to 10 days. In this note we focus mostly on the 6–10-day (averaged) range.

The 14 members of this ensemble can all be interpreted as starting from the same initial condition (IC), $t = 0$, with small perturbations superimposed so as to obtain 14 possible scenarios of future weather. The perturbations are either the automatic result of starting from time-lagged ICs (i.e., observations assimilated between the initial time of this member's integration and $t = 0$ cause the perturbation) or the result of adding deliberately to the analysis at $t = 0$ small perturbations, which, in their spatial structure, are optimal in some sense (Toth and Kalnay 1993). The configuration and nomenclature (member 1 and so on) of the 14 runs are represented schematically in Fig. 1, and for further discussion of this particular setup, the reader is referred to Tracton and Kalnay (1993).

As shown, for instance, in the results of the DERF90 experiment (Van den Dool 1994), individual model forecasts have on average some skill out to about 15 days, but already in the 6–10-day range skill is often quite small. It is for this reason that we need to change philosophy from single-run quasi-deterministic thinking toward a probabilistic approach that in all likelihood has to be based on multiple runs.

The 14 members of this ensemble are not "equal," because 1) they have different age at the target verification time (due to the time-lagging aspect), 2) they are produced at different resolutions (some at T126 truncated to T62 at day 6; others T62 from the beginning, and so on), and 3) some start from presumably the best possible IC (the analysis) while others start from a purposely perturbed IC. This inequality among the members leads us to ask how the members should be weighted, a question that has been formally addressed before (Dalcher et al. 1988).

This note is about the weights in constructing an ensemble average 6–10-day forecast at NMC. [Work on ensembles done elsewhere is described in Mureau et al. (1993).] While the creation of an ensemble average leads to a lot of interesting questions (which are the subject of this note), we point out that very likely there are many more questions to be asked about the use of ensembles, in particular as they relate to quantifying uncertainty. Here we stick to the seemingly simple task of making the best average.

In making an ensemble average it is very important to understand the averaging method used, its shortcomings, and the peculiarities of verification scores most commonly used to assess forecast accuracy/skill—that is, root-mean-square (rms) error and anomaly correlation (AC). Section 2 lists a number of properties of methods and verification scores and other considerations that are highly relevant to the issues of ensemble averaging and improving scores in general. In section 3 we present the particulars of the method

Corresponding author address: H. M. van den Dool, Prediction Branch, Climate Analysis Center, NOAA/NWS/NMC, Washington, DC 20233.

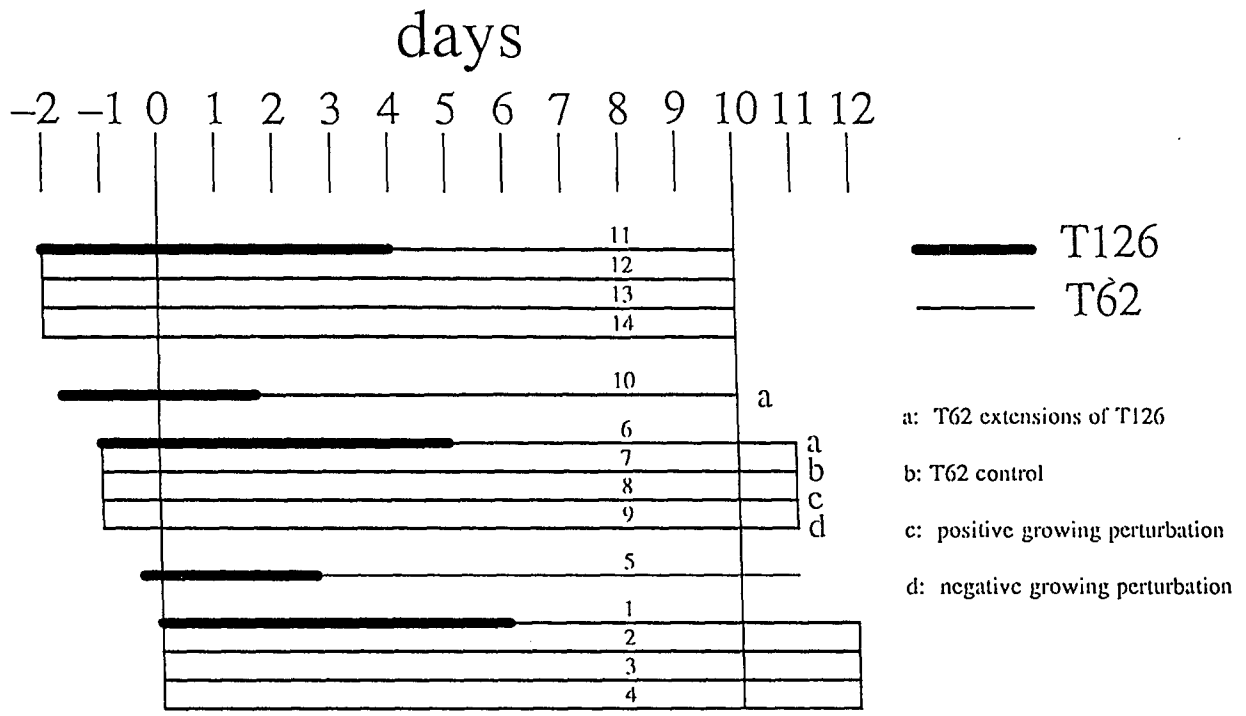

FIG. 1. A layout of the 14 integrations that constitute the ensemble.

used here, a version of multiple linear regression. The data are described in section 4 and the main results are presented in section 5. The paper is concluded with recommendations for application.

2. Considerations

a. About an ideal ensemble

An ideal ensemble may be thought of to consist of an infinity of "equal" members, equal in the sense that all members are plausible forecasts of the future. In that case their weights are a priori known and there is no need for any algorithm (always somewhat arbitrary) to determine the weights that would have given the best performance (by some arbitrary measure) over a training dataset.

Under the above ideal scenario, taking an ensemble average will remove all phenomena on which the members disagree completely and retain those phenomena that show up at the same time and place in all forecasts. This is the most ideal spatial filter imaginable, as it depends on the flow itself, changes from day to day, and does not depend on any a priori choices. Sometimes a cyclone will be predictable at day 5, sometimes it will not, and the ensemble-average "filter" allows for such temporal variations in the amount of filtering needed.

Obviously the ideal ensemble average has a reduced anomaly magnitude that has important repercussions for the rms error and AC. It is well known and fairly obvious that damping a forecast toward climatology reduces the rms error but up to a point only. But since the amount of damping (and its spectral distribution) is different every day, the averaging operator acting on an ideal ensemble is much more sophisticated than trivial overall damping, truncation, or other predetermined smoothing. To understand the impact of ensemble averaging on the AC, we here define

$$\mathrm{AC} = \frac{\sum\sum A'F'}{(\sum\sum A'A' \sum_t \sum_s F'F')^{1/2}} = \frac{\mathrm{cov}(A, F)}{sd(A)sd(F)}, \quad (1)$$

where A and F stand for "analysis" and "forecast," respectively, the summation is over time (t) and space (s), and the $'$ denotes a departure from climatology. An ideal ensemble average will leave the covariance unchanged, but the AC will increase nevertheless because $sd(F)$ decreases (relative to a single member). It is important to note that the observations are not touched by averaging some forecasts.

Below we discuss the real-world situation.

b. Formal procedures

There are formal procedures to minimize the rms error of a set of given forecasts, the procedure usually being based on linear regression, such as below in section 3. However, there is no formal procedure to maximize an (linear) AC, although users would also like to see a higher AC, aware as they are that part of the

reduction in rms error may be due to simple damping. The AC has become a widely accepted verification tool, but fundamentally, the meaning of an anomaly (or any) correlation is that the rms error of a set of raw forecasts could have been minimized if we had multiplied the (standardized) forecast anomalies by AC.

c. Ad hoc methods to increase AC

There are ad hoc methods to increase the AC, which at the same time are interesting diagnostics as to what part of the flow is best predicted/most predictable. Van den Dool and Saha (1990) defined an AC as a function of total or zonal wavenumber; that is, in (1) we sum (s) only over a certain set of wavenumbers. It is easy to see from their Figs. 2a,b that by truncating the forecast and/or analysis anomalies, so as to retain only the larger waves, the AC will go up. In this case both $sd(F)$ and $sd(A)$ go down, while the cov(A, F) goes down only very slightly. Retaining projections of forecasts/analysis onto leading empirical orthogonal functions (Branstator et al. 1993) falls in the same category. As explained before, the ensemble average is a (flow dependent) operator to retain the most predictable elements and therefore is almost certain to increase the AC by lowering $sd(F)$. Once more—taking an ensemble average does not change the verifying analyses, that is, does not lower $sd(A)$. In both Branstator et al. (1993) and Van den Dool and Saha (1990), the verification fields are filtered as well [lowering $sd(A)$]. We believe that users of forecasts are not necessarily helped by filtering analyses or observations.

d. What was done until now?

Until now the forecaster, based on his experience and the recent track record of various models, assigned weights to today's and yesterday's single-membered 6–10-day forecasts made by 1) the National Meteorological Center (NMC) and 2) the European Centre for Medium-Range Weather Forecasts, and constructed a "blend" out of it. (Note that single weights are applied to entire Northern Hemisphere maps. There are no space-dependent weights.) This is similar to an ad hoc ensemble average over at least 3 members. The more recent forecasts are usually weighed more heavily. More "members" are sometimes thrown in the blender by adding recently observed flows, persistence of the day 1–5 averaged forecast, conservative extension of the MRFs day 5 forecast out to day 10 using the model described in Van den Dool (1991), and so on.

e. Why do we want an ensemble average?

Taking an average is the simplest thing to do when having too many maps to look at. Also, the average could in some sense be considered the quasi-deterministic large-scale and low-frequency part of the forecast, and it fits in the long tradition of having just a single (or very few) forecasts at one's disposal. An average is also needed to calculate the spread of the members; that is, uncertainty is relative to some mean. Finally, the ensemble average is, under certain conditions, a probabilistic forecast expressed in a simple form. As long as probability forecasts are as simple as shifts of the unconditional probability distribution toward above or below the climatological mean, the specification of just the mean is in fact sufficient and complete probability information. Under less simple conditions, if the atmosphere were to bifurcate into a small number of discrete states, the ensemble average would be completely misleading probability information.

3. The method

Given daily a set of N 6–10-day average forecasts F_i, $i = 1, N$, available over a period of M days with associated 5-day mean verification (i.e., analyses A) available as well, we here ask the question of how to construct an ensemble average. We approach this, arbitrarily perhaps, through linear regression that will give weights such that the ensemble average has the smallest rms error over a training dataset. Assuming that the observed climatology has already been removed from all forecasts and analyses, we can define the ensemble average FE by

$$FE(s, t) = a_0 + \sum_{c=1}^{N} a_i F_i(s, t), \qquad (2)$$

where, as before, s is space (either a gridpoint index, or a spectral index) and t is time running from 1 to M. [There is no explicit reference to forecast lead time in (2). All forecasts F_i are 5-day averages valid at the same target time, 6–10 (8–12) days ahead for the youngest (oldest) forecast.] The $N + 1$ weights are to be determined from minimizing the residual Q given by

$$Q = \sum_t \sum_s [FE(s, t) - A(s, t)]^2. \qquad (3)$$

Differentiation with respect to $N + 1$ a_i's leads to the following equations:

$$a_0 + \sum_i a_i \bar{F}_i = \bar{A} \qquad (4a)$$

$$a_0 \bar{F}_j + \sum_i a_i \overline{F_i F_j} = \overline{A F_j}, \qquad (4b)$$

where the overbar stands for averaging over t and s. Using the first equation to eliminate a_0, the set of 14 in (4b) can be written in matrix form as

$$(\overline{F_i^* F_j^*}) \begin{pmatrix} a_1 \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} \overline{A^* F_1^*} \\ \vdots \\ \overline{A^* F_N^*} \end{pmatrix}, \qquad (5)$$

where the symbol * denotes a departure from the space–time mean. Equations (5) and (4a) can be solved by

standard means. The weights a_i are not a function of space at this point. The sum of the weights is not required to be unity.

On the training dataset the ensemble average according to (2) will have the lowest possible rms error. To judge improvement, we will also carry results for (i) individual forecasts, the latest high-resolution run (member 1) in particular; (ii) the straight arithmetic average over the 14 members; and (iii) a weighting scheme where the weights a_i are taken to be proportional to the expected anomaly correlation of the ith member. [The latter would be the result of setting all off-diagonal elements in the matrix in (5) equal to zero—that is, neglect all correlation between forecasts.]

4. Description of data

Our calculations are based on the ensembles of 14 member 6–10-day averaged 500-mb forecasts for the larger part of the Northern Hemisphere (20°–80°N) for the winter period: 14 December 1992 through 14 March 1993. So we had 87 cases (four forecasts were lost).

5. Results and discussion

Table 1 shows a correlation matrix for winter 1992/93. The correlation is not used explicitly in (5) but is informative and can be obtained from the covariances in (5) by dividing through by the appropriate standard deviations. Obviously, the diagonal becomes unity. The right-hand side column represents very nearly the usual ACs for each of the participating forecasts. With reference to Fig. 1, one can see that the AC generally decreases with age of the forecast, from 0.56 for the high-resolution most recent forecast to 0.40 for the oldest lower-resolution forecast. The (symmetric) 14 × 14 matrix on the left-hand side in Table 1 represents the correlation among the forecasts. It is worth reminding ourselves that we are not dealing with a perfect model, as is evident for instance from the fact that forecasts correlate higher with each other than with the verification. As it turns out the condition code of the matrix is very good, so that numerical problems are small and an inversion gives reliable a_i's. One reason for this is that forecasts do not correlate too highly with each other in the 6–10-day range as they would say at day 1.

Table 2 shows the weights a_i for winter 1992/93. The weights in Table 2 should not be taken too literally. For this sample size we estimate the uncertainty to be at least 0.05. The conclusions are as follows:

1) Forecasts older than 24 h (members 6–14) contribute little to the ensemble average. The ensemble size is nominally 14 but really more like 5 or even 4 as far as obtaining an ensemble average is concerned. This statement remains true even if the order 0.05 weights for the older members turn out to be statistically significantly different from zero on a larger dataset.

2) Four forecasts contribute almost equally, that is, the latest high-resolution run, the plus and minus perturbation T62 runs, and the 12-h-old T62 "aviation" run (member 5).

3) The ± perturbations (Toth and Kalnay 1993) are a helpful new element, carrying half the weight of the ensemble mean.

4) The bias error is about −20 gpm, which leads to $a0 = 13.6$ ($a0$ is smaller than 20 because the ensemble average has a reduced anomaly amplitude).

5) The T62 control forecast appears redundant (very low weight) in the context of having other fore-

TABLE 1. The anomaly correlation among the forecasts (14 × 14 matrix) and the anomaly correlation of each of the members with observed (column vector, size 14). All refer to 500-mb height 20°–80°N, averaged over days 6–10 for winter 1992/93. The numbering of the members (1–14) is as in Fig. 1, while the abbreviated decriptors HR, *c*, +, −, and *A* refer to high-resolution, control, positive and negative perturbation, and aviation run, respectively.

	HR	*c*	+	−	*A*	HR	*c*	+	−	*A*	HR	*c*	+	−	OBS
1	1.00	0.76	0.70	0.71	0.76	0.67	0.62	0.59	0.59	0.60	0.56	0.52	0.49	0.52	0.56
2	0.76	1.00	0.83	0.84	0.69	0.65	0.71	0.66	0.65	0.59	0.54	0.58	0.55	0.56	0.56
3	0.70	0.83	1.00	0.70	0.64	0.62	0.68	0.70	0.60	0.56	0.53	0.56	0.56	0.54	0.54
4	0.71	0.84	0.70	1.00	0.67	0.63	0.66	0.60	0.67	0.58	0.53	0.57	0.51	0.55	0.55
5	0.76	0.69	0.64	0.67	1.00	0.74	0.64	0.60	0.61	0.67	0.58	0.54	0.49	0.52	0.54
6	0.67	0.65	0.62	0.63	0.74	1.00	0.69	0.63	0.65	0.70	0.62	0.56	0.52	0.53	0.47
7	0.62	0.71	0.68	0.66	0.64	0.69	1.00	0.77	0.79	0.62	0.59	0.66	0.60	0.60	0.47
8	0.59	0.66	0.70	0.60	0.60	0.63	0.77	1.00	0.64	0.57	0.56	0.62	0.65	0.55	0.46
9	0.59	0.65	0.60	0.67	0.61	0.65	0.79	0.64	1.00	0.60	0.58	0.60	0.53	0.61	0.47
10	0.60	0.59	0.56	0.58	0.67	0.70	0.62	0.57	0.60	1.00	0.69	0.59	0.54	0.55	0.46
11	0.56	0.54	0.53	0.53	0.58	0.62	0.59	0.56	0.58	0.69	1.00	0.63	0.57	0.58	0.39
12	0.52	0.58	0.56	0.57	0.54	0.56	0.66	0.62	0.60	0.59	0.63	1.00	0.71	0.74	0.40
13	0.49	0.55	0.56	0.51	0.49	0.52	0.60	0.65	0.53	0.54	0.57	0.71	1.00	0.58	0.38
14	0.52	0.56	0.54	0.55	0.52	0.53	0.60	0.55	0.61	0.55	0.58	0.74	0.58	1.00	0.40
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	

TABLE 2. Optimal weights for members 1–14.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	a0
0.15	0.03	0.16	0.14	0.15	−0.05	−0.03	0.04	0.07	0.08	−0.04	−0.03	0.04	0.05	13.6

casts available. One reason could be that the control run has nothing to add to what has already been contributed by the average of the + and − perturbation runs. By current construction (Toth and Kalnay 1993) the average of + and − equals the control at $t = 0$, and for as long as a linear assumption remains valid there is redundancy. We checked this explanation by rerunning a 3×3 matrix problem involving only members 2, 3, and 4. In that setting the weights are 0.14(c), 0.25(+), and 0.25(−). Thus, the explanation is partly true and it would therefore be more effective to change the setup such that the + and − are not exactly symmetric relative to the control. The further lowering of the weight to near zero (for T62 control) when 14 members are admitted points to additional redundancies of the T62 control relative to other forecasts or linear combinations thereof.

The weights have only a vague similarity to the ACs (rhs in Table 1). In other words, the weights are not typically proportional to the AC and do not decrease with age like the ACs on the right-hand side of Table 1. This, of course, is the whole idea of optimal averaging, which takes advantage of the correlations among the forecasts, or, alternatively, of the partial correlations hidden underneath the matrix in Table 1.

We reran the problem with only members 1, 3, 4, and 5 and found very similar weights: that is, assuming zero weight for 10 members did not change materially the relative weighting of the 4 most important forecasts.

Note in Table 2 that the sum of the absolute weights is slightly above unity (1.06). Therefore, there is no explicit mathematically imposed damping. Any reduction in anomaly amplitude is the result of applying the ensemble averaging procedure and in doing so filtering the less predictable scales, as it should.

Table 3 shows the scores. By taking a straight average over the 14 members there is already a noticeable improvement over the latest high-resolution run, both by rms (98–87 gpm) and AC (0.560–0.589) standards. A similar substantial improvement can be obtained by the optimal averaging (rms down to 79 and AC up to 0.623). In between but close to the arithmetic mean is an averaging procedure where the weights were proportional to the anticipated AC (constant over the whole period). Clearly, knowing the correlation among the forecasts is very helpful. In all we seem to have improved the forecast by about six AC points and the rms is down from 98 to 79 gpm.

The last column of Table 3 shows the anomaly magnitude (defined as the square root of the space mean-squared anomalies) of the forecasts. Clearly, the optimally averaged forecast has a small anomaly amplitude (69), both compared to member 1 (101) and the arithmetic mean ensemble average (85). The optimal average removes the least predictable parts of the flow more effectively and completely than any other averages.

When we added as a 15th member the day 1–5 averaged forecast averaged over members 1, 3, 4, and 5, the rms and AC improved further to 78 and 0.640, respectively (see Table 3, entry denoted "Optimal 15"). Table 4 shows the weights for this case. Some conservatism helps—member 15 has the highest weight.

We made one attempt to derive the regression as a function of total wavenumber. In (2) the weights a_i do not depend on location or spectral scale. A practical disadvantage of regression by wavenumber is that sampling uncertainty would be a bigger problem. The weights are shown in Table 5 when derived for three separate total wavenumber bands. For each band we find weights broadly similar to those for all waves together. Therefore, there is no great need to make the regression wavenumber dependent. This result appears at first sight different from Dalcher et al. (1988). In their case, forecasts were available only at daily intervals (we have 5 per 24-h interval), so their ensemble average was primarily the most recent forecast, with older forecasts coming in with very small weights only. On a set of M single forecasts, a regression acts as an overall damping of anomaly magnitude so as to minimize the rms error. When offered a regression as a function of total wavenumber, the damping will be scale dependent (Dalcher et al.'s case). This is already somewhat accomplished in our case by averaging four forecasts with nearly equal weight. From an rms error minimization standpoint, the best damping is achieved when the am-

TABLE 3. Root-mean-square error and AC for different strategies.

	Root-mean-square	AC	Anomaly magnitude
Latest run	98	0.560	101
Average 14	87	0.589	85
AC-avg 14	86	0.596	86
Optimal 14	79	0.623	69
Optimal 15	78	0.640	71
Poor man	84	0.570	70
Days entered individually	78	0.642	69

TABLE 4. Optimal weights with member 15 added.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	a0
0.13	0.04	0.13	0.13	0.13	−0.05	−0.02	0.03	0.07	0.07	−0.04	−0.03	0.03	0.05	0.18	14.0

plitude for each wave is reduced by a factor proportional to the AC for that wave.

Another way of discussing the issue of damping the forecast is therefore by showing the spectrum of the latest MRF run in comparison to that of the optimally averaged ensemble. We shall look upon the latter as a filtered version of the first. Figure 2 shows the amount of damping that results, on average, from taking an ensemble average—more precisely, the ratio of the anomaly amplitude after filtering to that before, as a function of wavenumber. Consistent with Table 3, overall the amplitudes are reduced to about 70%. A good deal of scale-dependent damping is accomplished with essentially 4 members, the lower (higher) wavenumber being reduced to 80% (60%). We suspect that somewhat stronger scale dependence is required such that the curves in Fig. 2 resemble more closely the AC as a function of n and m, as shown, for example, in Van den Dool and Saha (1990, Fig. 2). We hope to accomplish this in the near future by larger ensemble sizes (about 40 members instead of 14 have been implemented in early 1994).

A poor man could approximate the filtering effect of the ensemble average by applying the average damping presented above every day to the latest MRF run. We did this as a test to make sure the improvements in scores reported earlier cannot be obtained in a trivial way. Numbers under the entry in Table 3 labeled "poor man" indicate that, indeed, the optimally weighted ensemble average (labeled optimal 14) is much better than the poor man's proxy. Apparently, plain and constant (in time) damping explains little or nothing of the gains reported before. Mostly, the poor man misses out on very substantial time variations in the amount and spectral distribution of the damping to be applied, while having an ensemble is a means to anticipate the dispersion among the ensemble members. High (low) dispersion corresponds to strong (weak) filtering.

As can be seen from the value of $a0$ in Table 2, the forecasts suffer from a cold bias. The ensemble average forecast corrects for this bias, although only in a spatial mean sense. Operationally at NMC forecasts are postprocessed and a space-dependent correction is applied based on the mean forecast error in the last two months (Alpert and Saha 1989). In the future operational setup, it would be best to first postprocess each forecast F_i separately in this manner before regression (2) is applied.

So far we have used 6–10-day time-averaged forecasts to create an ensemble average that matches 6–10-day averaged analyses as closely as possible. One could question the wisdom of time averaging both on the predictand and predictor side. (In some sense time averaging forecasts is a poor man's ensemble average.) As an additional experiment we entered forecasts for days 6, 7, 8, 9, and 10 individually to predict the 6–10-day average. Using all 14 members, this leads to a 70 × 70 matrix in (5), which we solved. We found, however, that we can retain the essence of this exercise by using the first 5 members only, thus solving a 25 × 25 matrix. The remaining 8 members contribute

TABLE 5. Optimal weights for various total wavenumber bands.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	a0
All waves														
0.15	0.03	0.16	0.14	0.15	−0.05	−0.03	0.04	0.07	0.08	−0.04	−0.03	0.04	0.05	13.6
waves 0–6														
0.16	0.02	0.17	0.22	0.18	−0.11	−0.04	0.02	0.10	0.10	−0.08	−0.06	0.04	0.07	15.0
waves 7–12														
0.16	0.06	0.11	0.10	0.16	−0.12	0.02	0.03	0.04	0.07	−0.01	0.01	0.03	0.04	−0.8
waves 13–30														
0.08	0.05	0.06	0.05	0.09	0.01	0.02	0.02	0.03	0.05	0.00	0.03	0.01	0.03	−0.8

little, neither when entered as individual days nor, as shown before, when entered as 5-day averages.

Table 6 shows the weights for this experiment. Estimating 25 weights (rather than 14 as before) introduces a more sample-specific result. We also note that earlier attempts to estimate the weights of individual days using a single member (member 1) encountered some difficulty in terms of obtaining numerically stable weights. The matrix used here had a favorable condition code. The results are largely consistent with those discussed before in terms of the relative weights of 6–10-day averaged members 1 to 5—that is, the control T62 forecast contributes little. However, we also note that in general day 6 contributes most, with days 7, 8, 9, and 10 coming in with much less weight. Recalling how much the $D + 3$ (member 15) contributed (see Table 3 and 4), this reconfirms and extends an earlier finding by Saha and Van den Dool (1988) that at some point into the forecast a continued integration yields little or no new information. The verification of forecasts based on individual days and members 1–5 only is shown at the bottom in Table 3. Gains from using individual days as predictors are negligible, even on dependent data. Therefore, use of time-averaged predictors is justifiable and saves time.

6. Conclusions and recommendation

We believe we have demonstrated that an optimal ensemble average is considerably better than a single forecast for the average of day 6–10. Scores in terms of rms and AC for Northern Hemisphere 500-mb height are substantially better when an ensemble average is used. This gain is nontrivial and cannot be reproduced from "just" the overall or scale-dependent damping effect associated with taking an ensemble average. The real gain of having an ensemble is associated with day-to-day variation in the amount of dispersion among the members of the ensemble, which implies day-to-day variation in the amount and spectral distribution of the damping associated with taking the mean.

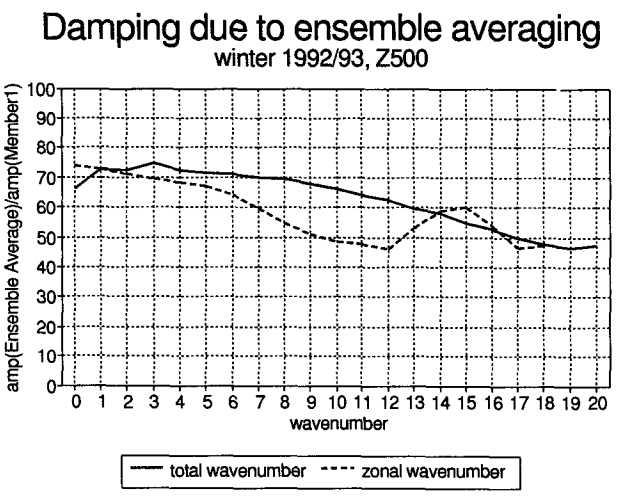

FIG. 2. The ratio of the amplitude of anomalies in the optimally weighted ensemble mean to those in the latest single high-resolution run, as a function of zonal (dashed line) and total (solid line) wavenumber. The damping represents an average over the whole dataset of winter 1992/93.

TABLE 6. Weights for individual days of members 1–5.

Member/Day	1	2	3	4	5	Sum
6	0.14	0.01	0.09	0.08	0.03	0.35
7	0.03	−0.01	0.02	0.01	0.03	0.08
8	0.03	0.01	0.06	0.04	0.03	0.17
9	−0.02	0.02	0.03	0.03	0.02	0.08
10	0.02	−0.01	−0.01	0.03	0.04	0.07
Sum	0.20	0.02	0.19	0.19	0.15	0.75

As far as accuracy of the ensemble mean is concerned, only the recent members contribute substantially. These are the latest high-resolution run (member 1), the ± perturbed 0000 UTC run, and the 12-h older aviation run (member 5). The T62 control run contributes little or nothing to the ensemble mean.

The fact that only 4 members contribute significantly to the accuracy of the ensemble mean should not be construed to mean that the size of the ensemble need not be larger than 4. The conclusion applies only to this particular setup, which is a compromise between having as many members as possible from the most recent initial time and the practicalities of runs in real time with deadlines for operational usage. It is no accident that the older members appear to contribute little. We speculate that more forecasts (if properly perturbed) from the most recent initial time(s), as planned to be implemented as of February 1994, would help. Also we here discuss only the accuracy of the mean, and our conclusion may not carry over simply to usage of ensembles in general. Research by Tracton appears to indicate that probability forecasts do benefit from the older members in the ensemble.

For operational application the following is reasonable. Weights will be based on the last two months and will be updated once a week. Two months is a compromise between changing seasons and the need for a long sample to arrive at reliable weights. This application has already found its way into various NMC products since mid-1993, most notably into the official 6–10-day averaged 500-mb height anomaly map.

Preliminary tests of the above for spring 1993 show that most of the gain in skill survives the test on independent data. Basically, this is because we estimate only about four coefficients from a very large dataset (60 time levels, Northern Hemisphere area).

If possible, the forecasts should be bias corrected before they enter (4a)–(5). We can further include the ECMWF forecast, or any in-house forecasts such as

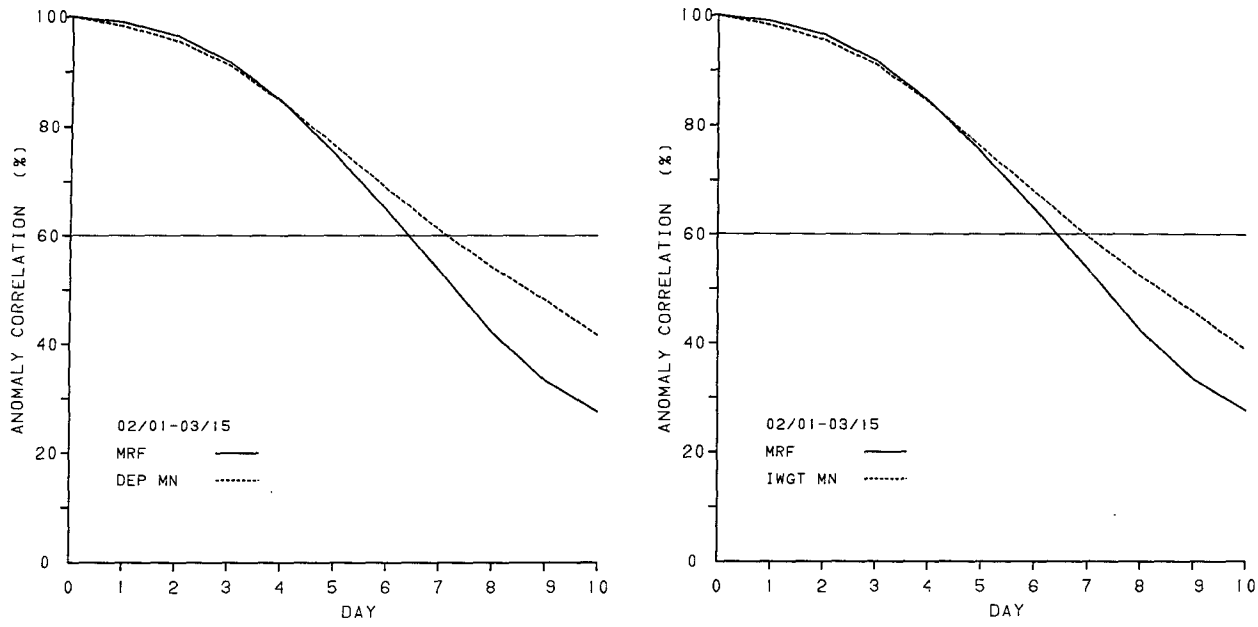

FIG. 3. The anomaly correlation of daily 500-mb height forecasts for leads of 1–10 days on the 20°–80°N Northern Hemisphere domain for the period 1 February 1993–15 March 1994. The solid line is for the latest MRF run, and the dashed is for the optimal ensemble average. On the left are the dependent weights and on the right are the independent weights estimated from verifiable forecasts during December 1992 and January 1993.

the anomaly vorticity advection model, so as to guide the 6–10-day forecaster objectively.

The ± perturbations in the T62 model were designed to achieve maximal divergence between the control and perturbed runs. Have we succeeded in this? Returning to Table 1, the correlation among the four 0000 UTC runs is given in the upper left. It appears that the T62 control diverges less from the + (or −) runs than it does from any other forecast. Indeed the 0.83 and 0.84 correlations are by far the highest in the whole matrix. This has to be studied more, keeping in mind that both the structure of the ± perturbation and their initial magnitude may explain the high correlations.

The method, as is, can be applied to any parameter, any lead time, and any level. We anticipate weights somewhat like Table 2 but not necessarily very close. At NMC the operational 6–10-day forecast is naturally focused on an area smaller than 20°–80°N (all longitudes) for which we did the calculations. It is worthwhile to investigate the method on a smaller domain, although the user better be aware of the reduction in sample size. Also there is no obvious a priori reason why the weight should depend on the area (other than for sampling reasons).

In addition to the good contributions from the members that are perturbed by the modes described in Toth and Kalnay (1994), we are pleasantly surprised by the performance of member 5, the so-called aviation run. This integration (member 5) starts from 1200 UTC 12 h earlier, and yet contributes to the ensemble average on par with three integrations from 0000 UTC. One can see in Table 1 that the aviation run on its own scores as highly as the 0000 UTC runs. Apparently, the 1200 UTC initial condition contains information that cannot be accounted for by perturbing the 0000 UTC initial condition. We can only speculate why this is: the daily cycle? The availability of data? One must keep this in mind also when comparing forecasts from 0000 UTC (many at NMC) to those from 1200 UTC (the 10-day EC forecasts).

Given a 14-member ensemble, we determined 14 weights. There is no a priori reason why the positively and negatively perturbed members should have different weights. As shown in Tables 2, 3, and 5, the weights for members (3, 4), (8, 9), and (13, 14) were indeed very close, and we declare their difference a matter of sampling. To further suppress sampling error, one could rewrite the problem, assuming upfront $a_3 = a_4$ etc., which reduces the matrix to 11×11. We may do this in the practical application in the future, after we settle on a configuration that will be used for a long time.

For further demonstration, we applied the method to the individual daily 500-mb height forecasts from day 1 to 10. The results were provided by J.-F. Pan. Figure 3 shows on the left a comparison of the AC for member 1 (full) and the optimal average (dashed) for leads of 1–10 days, for the period 1 February 1993–15 March 1994. This graph is a powerful demonstration of the improvement of the scores because using ensemble averages amounts to a gain of about half a day at the 0.6 AC level, a gain normally associated with years

of hard work on model improvements, increased resolution, or new observational platforms. The right portion of Fig. 3 shows the same but with weights determined from the verifiable forecasts in December and January. The results appear thus quite satisfactory on independent data. Early on in the forecast (at day 1) the correlation among the members' 500-mb height forecasts is extremely high (0.99), thus making the ensemble approach not only less relevant but also mathematically difficult. A stable solution for the weights at short forecast leads was obtained by applying a modest amount of "ridging" as described in Meissner (1978). Clearly, the beneficial effects of optimal ensemble averaging (relative to using the most recent high-resolution run) do not show until the correlation of the members with each other has decreased to 0.75 or so. For heights this is at about day 5.

Although the results so far are encouraging for 500-mb height forecasts it should be pointed out that, as of now, no hard evidence has been presented that the forecast of surface weather elements has improved. This will be studied further at NMC.

Acknowledgments. The authors gratefully acknowledge the fruitful discussions and feedback from M. S. Tracton, J.-F. Pan, R. Martin, E. Kalnay, Z. Toth, R. E. Livezey, D. R. Rodenhuis, and E. Berry.

REFERENCES

Alpert, J., and S. Saha, 1989: Operational systematic error correction for the NMC operational MRF model. NMC Office Note 360.

Branstator, G. A., A. Mai, and D. Baumhefner, 1993: Identification of highly predictable flow elements for spatial filtering of medium- and extended-range numerical forecasts. *Mon. Wea. Rev.*, **121,** 1786–1802.

Dalcher, A., E. K. Kalnay, and R. N. Hoffmann, 1988: Medium range lagged average forecasts. *Mon. Wea. Rev.*, **116,** 402–416.

Meissner, B. N., 1979: Ridge regression. Time extrapolation applied to Hawaiian rainfall normals. *J. Appl. Meteor.*, **18,** 904–913.

Mureau, R., F. Molteni, and T. N. Palmer, 1993: Ensemble prediction using dynamically conditioned perturbations. *Quart. J. Roy. Meteor. Soc.*, **119,** 299–324.

Saha, S., and H. M. van den Dool, 1988: A measure of the practical limit of predictability. *Mon. Wea. Rev.*, **116,** 2522–2526.

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

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

Van den Dool, H. M., 1991: Mirror images of atmospheric flow. *Mon. Wea. Rev.*, **119,** 2095–2106.

——, 1994: Long range forecasts through numerical and empirical methods. *Dyn. Atmos. Oceans,* **20** (3), 247–270.

——, and S. Saha, 1990: Frequency dependence in forecast skill. *Mon. Wea. Rev.*, **118,** 128–137.