The effect of increased horizontal resolution on the NCEP global ensemble mean forecasts

I. Szunyogh[1]

UCAR Visiting Scientist at NCEP, Camp Springs, Maryland, USA;

Z. Toth,

GSC, NCEP, Camp Springs, Maryland, USA

submitted to *Monthly Weather Review*, October 2000

June 4, 2001

[1] *Corresponding author:* University of Maryland, Institute for Physical Science and Technology, Computer and Space Sciences Building, College Park, MD, 20742-2431, E-mail:szunyogh@ipst.umd.edu

Abstract

The primary goal of this paper is to explore why the use of increased horizontal resolution enhances the performance of the National Centers for Environmental Prediction (NCEP) global ensemble mean forecasts. Numerical experiments were carried out with a 10-member (5-pair) 0000 UTC subset of the NCEP global ensemble forecasts for a 30-day period during January-February 1999. Four sets of ensembles and corresponding control forecasts were generated. One ensemble was identical to the then operational T62 horizontal resolution NCEP ensemble, while in the other three ensembles the horizontal resolution was increased to T126 out to 1-, 3-, and 15-day forecast lead times.

Anomaly correlation and root-mean-square error (also decomposed into bias and variance terms) were used to evaluate the control and ensemble mean forecasts. As expected, the use of a higher resolution model improves both scores. A newly developed condition for optimal smoothing indicates that the root-mean-square error for the high resolution 10-member ensemble is nearly as low as it can be given its anomaly correlation. Therefore, further significant improvements in the ensemble mean forecasts can be achieved only through improved anomaly forecast patterns, and not through additional smoothing.

The two main meteorological aspects of the higher resolution induced error reduction for both the control and the ensemble mean forecasts are (1) the maintenance of a more realistic time-mean flow; and (2) the better prediction of high frequency transients along the mid-latitude storm tracks. The effect of increased horizontal resolution, however, is markedly more positive on the ensemble mean than on the control forecasts. This is because the ensemble mean (1) efficiently filters out unpredicatble small scale features at high resolution; and (2) accentuates the relatively large systematic errors present in the low resolution integrations.

1 Introduction

The long time limit of the root-mean-square (rms) error is $\sqrt{2}$ times larger for a numerical weather forecast than for a prediction based on climatology. This happens because a single numerical model forecast provides the same level of details at all lead times, irrespective of the various predictability time limits associated with different weather phenomena. Therefore, single forecasts, correctly predicting some deviations from the climate, become aggravated by unpredictable features which leads to a rapid degradation of forecast quality in terms of rms error. Ensemble forecasting was introduced partly to remedy this problem. As Leith (1974) pointed out *ensemble averaging is a potentially optimal nonlinear filter* since the ensemble mean would be the best unbiased estimate of the true state of the atmosphere in an rms sense if (1) the model was identical with the real atmosphere; (2) the climate attractor was ergodic (i.e. the climatological phase-space and time averages were equal), and (3) the ensemble was perfect (i.e., it had infinite number of members and the members represented equally likely states of the atmosphere). In fact, several papers documented that the mean of an ensemble of numerical weather forecasts becomes clearly superior to a single control forecast as forecast lead time increases, even under realistic conditions when imperfect model and ensemble formulation is used (e.g. Toth and Kalnay, 1993, Houtekamer and Derome, 1995, Molteni et al, 1996).

The main goal of this paper is to explore *why* the use of increased horizontal resolution enhances the performance of the ensemble as a *nonlinear filter*. This research was motivated by the finding, also presented here, that the increased horizontal resolution has a much more positive effect on the performance of the ensemble mean than on the quality of the single deterministic forecasts. Our work started in 1999, when upgrades to the global Ensemble Forcasting System (EFS) could be considered following the acquisition of a new Class-VIII parallel supercomputer at the National Center for Environmental Prediction (NCEP). Encouraged by studies at the European Centre for Medium-Range Weather Forecasts (ECMWF), which revealed the relative importance of adequate model resolution for their ensemble predic-

tion system (Buizza et al., 1998) , a possible increase in horizontal resolution was considered first.

Since ensemble averaging improves the forecast scores by smoothing the meteorological fields, a thorough assessment of the quality of an EFS should also consider whether this filtering effect properly reflects in an inverse fashion the level of predictability that typically decreases with increasing lead time. Leith (1974) and Houtekamer and Derome (1995) compared the performance of the ensemble mean and a forecast that was empirically smoothed based on forecast error history information. In this paper, we present an alternative approach to assess whether the smoothing effect of an ensemble is optimal. This includes the derivation of a condition for optimal smoothing and the decomposition of the rms error into bias and forecast error variance terms. The latter technique is routinely used to monitor and analyze deterministic numerical forecasts at NCEP (White, 1999), but it has not been used with ensembles. We will demonstrate that separating the bias component of the rms error can be rather instructive since this term quantifies the difference between our model and reality in a time mean sense. Apparently, ensemble averaging cannot be expected to remove this part of the error if all members of the ensemble are generated by the same model.

The structure of the paper is as follows. Section 2 details the set-up of the numerical experiments. Section 3 describes the theoretical relationships between rms error, anomaly correlation and the rms distance between forecasts and climatology, followed by related quantitative results for the control and mean forecasts at various resolutions. (Verification results based on probabilistic verification scores, like the Brier and the ranked probability scores (Wilks and Hamill, 1995, Talagrand et al., 1999, and Richardson, 2000), will be presented in a follow-up study). The decomposition of the mean square error into bias and forecast error variance terms, with associated results for the different forecasts are presented in section 4. Section 5 offers a discussion of the verification results, while section 6 presents the conclusions.

2 Experimental set-up

2.1 The sample time period

All experiments were carried out with a 10-member (5-pair), 0000 UTC subset of the NCEP global ensemble forecasts. Error statistics were accumulated for a 30-day period from January 13 through February 11, 1999. Though the choice of a contingent time period has the disadvantage of possibly having strong auto-correlations between errors of consecutive days, thus limiting the effective size of the statistical sample, it has the advantage that persistent, slowly varying error patterns can be detected in the ensemble.

This particular 30 day period, overlapping with the 1999 Winter Storm Reconnaissance program, was selected because a large number of diagnostics prepared for an earlier study (Szunyogh et al., 2000) were already available. Here, we show only eddy statistics that are crucial to exploring the relationship between the location of the storm tracks and the geographical distribution of forecast error reductions due to increased horizontal resolution. The eddy quantities are defined by the deviation from the monthly mean and the 500 hPa geopotential height variance, the meridional temperature flux and the vertical temperature flux are shown in Fig. 1. Overlapping regions of pole-ward and upward temperature fluxes mark areas of available potential to eddy kinetic energy conversion. The three main regions of baroclinic energy conversions (North Pacific, North Atlantic, and SH mid-latitudes) are well distinguishable. The seasonal differences are also well marked by the more intense temperature fluxes in the Northern Hemisphere. The propagation of the baroclinic wave packets in the Northern Hemisphere was also well documented for the sample period by Fig. 4 in Szunyogh et al. (2000).

2.2 The operational ensemble configuration

Until June 2000, the implementation of the operational global EFS at NCEP consisted of 17 16-day forecasts. All forecasts were made with the NCEP Medium Range Forecast (MRF)

model (Derber et al., 1998). In addition to the ten and four perturbed integrations made with a T62 horizontal resolution 28 vertical levels version of the model at 0000 and 1200 UTC respectively, two control integrations were made at 0000 UTC, and one at 1200 UTC. The high resolution control forecasts were run at T126 (T170 from June 15 through October 5 1998, and after January 24 2000) resolution with 28 (42) vertical levels out to 7- and 3-day (3.5-day) days lead time at 0000 and 1200 UTC respectively, after which the fields got truncated to T62 resolution and the runs were extended to 16-day lead time (Tracton and Kalnay, 1993, Toth and Kalnay, 1997)

While ECMWF uses leading initial and evolved singular vectors defined by a norm with energy dimension (Molteni et al., 1996 and Buizza et al., 1999), NCEP runs breeding cycles (Toth and Kalnay, 1993 and 1997, and Iyengar et al., 1996) to generate the perturbed initial conditions. In a breeding cycle, first half of the difference between the initially oppositely perturbed pairs of 24-hour forecasts is taken. This three dimensional difference field is then rescaled by a two-dimensional regional rescaling factor, $r(\lambda, \phi, t)$, which is varying with the geographical location, but is fixed for all model variables at all levels. The rescaling factor is defined by the ratio $r(\lambda, \phi, t) = mask(\lambda, \phi, t)/K(\lambda, \phi, t)$. Here, mask denotes the daily varying average RMS difference between two independently run analysis cycles, computed using the rotational kinetic energy at the 500 hPa pressure level as inner product and $K(\lambda, \phi, t)$ is the square root of the rotational kinetic energy at the 500 hPa pressure level for the difference field to be rescaled. Both fields are smoothed by a Gaussian filter before the ratio is computed, removing most of the variance in the kinetic energy field associated with wave-numbers larger than 8-9. The purpose of this rescaling procedure is to ensure that the initial ensemble perturbations are representative of the typical large-scale geographical distribution of the analysis uncertainty. The only reason why the large scale distribution of the rotational kinetic energy of the initial perturbations can depart from the mask at the 500 hPa level is that the size of the initial perturbations is never increased during rescaling: the value of r is set to 1 at locations where it would be larger otherwise.

The regional rescaling algorithm was designed with the aim of retaining the structure of synoptic and smaller scale features developed in the 24-hour evolved perturbations. Thus we can expect that by increasing the horizontal resolution the initial structure of the bred perturbations will also change since the magnitude and the structure of the 24-hour evolved perturbations may be different in the different resolution models and only the largest scale components of these perturbations are constrained strongly by the regional rescaling.

2.3 The experimental ensemble configurations

Four sets of ensembles were generated. The first ensemble (referred to as T62 hereafter) was an almost exact replica of the 0000 UTC low-resolution subset of the operational ensemble, while in the other three ensembles (referred to as D1, D3 and T126 hereafter) the horizontal resolution was increased to T126 out to 1-, 3-, and 15-day forecast lead times. To save computer time the D1 (D3) runs were stopped after 4 (7) days of model integration. A control forecast, started from the unperturbed analysis, was also run for each ensemble following the same truncation strategy that was applied to the associated ensemble.

Testing ensemble configurations that changed their resolution after 1 or 3 days was dictated by the operational constraint that an ensemble integrated at resolution T126 up to 16 days would not be affordable at NCEP in the near future. We note, however, that reducing the horizontal resolution of high resolution MRF control forecasts after the first few days of model integration has been a long time practice at NCEP. This strategy is based on the experience that increased horizontal resolution for the first few days of model integration has significant positive impact on forecast quality for the entire 16-day forecast range; a reduction of the horizontal resolution after a few days does not degrade the skill scores substantially. The general belief is that this is due to to the better quality of the higher resolution analysis and the shorter predictability time limit of the smaller scale weather phenomena. For example, experience shows that an analysis taken from a T126 cycle and truncated to resolution T62 usually leads to a better T62 forecast than the one that was started from an analysis of a T62

cycle (S. Tracton, 1993, personal communication). Because of this, the T62 perturbed initial analyses of the operational NCEP EFS have been created around a T126 analysis truncated to resolution T62 (Tracton and Kalnay, 1993).

It must be emphasized that reducing resolution after a few days of model integration is more than a simple truncation (spectral filtering) of the meteorological fields. It also means using a different model after the truncation. Most importantly, the reduced resolution model at the bottom of the atmosphere is forced by boundary conditions/orography that are different from those used in the high resolution model. Secondly, the physical parameterization schemes may behave differently in the different horizontal resolution models. Finally, while at T62 resolution the total wave-number of the smallest retained feature is 62, not all interactions between structures with wave-numbers smaller than 62 are retained (Machenauer, 1991 and Kadar et al., 1998). In fact, all of those interactions are neglected which would result in entities characterized by wave-numbers larger than 62. Consequently, the interactions between features characterized by total wave-numbers smaller than 62 are better represented, especially for the high wave-number components close to the cut-off wave-number 62, in a T126 version of the model.

All breeding cycles were initiated with the operational bred perturbations from January 3 1999. Since these perturbations had horizontal resolution T62 the spectral coefficients related to higher wave-numbers in the T126 runs were simply set to zero. Initial conditions were created by adding the perturbations to the operational T126 analysis of NCEP (truncated to T62) in the initially high resolution (low resolution) ensembles.

For running breeding cycles, a T126 version of the mask was also needed. This was prepared by first transforming the mask to the T62 spectral space from the related Gaussian grid, then filling up the spectral coefficients related to wave-numbers higher than 62 by zeros, and transforming the field to the high resolution Gaussian grid. As it was expected, after 3-4 days of running the breeding cycle at resolution T126 no transient behavior could be observed in the initially high resolution cycles.

After interpolating the forecasts and analyses to a $2.5^o \times 2.5^0$ grid, all ensemble and control forecast products were verified against the control analysis. We note that the resolution of this verification grid is lower than the resolution of the T62 analysis-forecast fields, which means that only features resolved in both the low and the high resolution runs are verified and the effect of high wave-number modes in the T126 simulations are taken into account only implicitly.

Verification results are presented for the 500 hPa geopotential height in the Northern Hemisphere (NH) and Southern Hemisphere (SH) mid-latitude regions. The NH (SH) verification region is defined by the 30N-70N (30S-70S) latitude band.

Some of the verification statistics require the knowledge of climatology at the grid-points on a daily basis. In this study, the computation of climatology (denoted by c), is done in two steps: first, the monthly averages of 36 years of NCEP reanalyses (Kalnay et al., 1996) are taken and then the daily values are computed by a linear time interpolation assuming that the monthly average is representative at the middle of a given month.

3 Rms error, anomaly correlation, and ensemble smoothing

3.1 Condition for optimal smoothing

The most widely used forecast score for the verification of the ensemble mean is the *rms error (RMS)* defined as the root-mean-square distance between the forecast and the verifying data set

$$RMS = \sqrt{< (f-a)^2 >}. \tag{1}$$

Here, f is the forecast and a is the verifying analysis, both given on the $2.5^o \times 2.5^0$ grid. The angled brackets stand for the mean that will be taken over the 30-day sample period and over all points within the verification domain. When definition of a local error is needed, the mean

is taken over the 30-day sample period at each grid point. This time-mean is denoted by an overbar, for instance, the *local rms error* is

$$RMS(\lambda, \phi) = \sqrt{\overline{(f-a)^2}}. \tag{2}$$

The *mean-square error (MS)*, the square of RMS, can be decomposed (Simmons et al., 1995) as

$$MS = < (f-a)^2 > = < (f-c)^2 > + < (a-c)^2 > -2 < (f-c)(a-c) > . \tag{3}$$

When c is computed on a daily basis and averages are taken for large regions over an extensive time period, $< (f-c) >$ and $< (a-c) >$ become negligible. In this sense, the first (second) term of the rhs. in Eq. 3 is usually referred to as forecast (analysis) variance. Here, especially when only the time averages are taken, $< (f-c) >$ ($< (a-c) >$) is not negligible and for clarity the mean-square distance between the forecasts (analyses) and the climatology will be referred to as the *mean-square forecast (analyzed) anomaly*.

For a good NWP analysis-forecast system the mean-square analyzed anomaly provides a good estimate of the climate variance for the selected sample period and the mean-square forecast anomaly (first term of rhs. in Eq.3) is near to the mean-square analyzed anomaly (second term of rhs. in Eq.3) for all forecast lead times. Since the covariance between the analyzed and the forecast anomalies (third term of rhs. in Eq.3) goes to zero as the forecast lead time increases, the long-time limit of MS (RMS) is twice ($\sqrt{2}$)as large as the (root-)mean-square analyzed anomaly. When climatology is used as forecast (f=c) the first and the third term on the rhs. in Eq.3 are identically zero. This means that the long time limit of MS (RMS) is reduced to the (root-)mean-square analyzed anomaly, but at the price of significantly increasing the short-term forecast errors. The use of an appropriate ensemble has the benefit of retaining the forecast anomaly $(f-c)$ at short lead times, similarly to a single forecast, and removing all anomalies, like a climatology based forecast, in the long time limit.

Operational EFS systems are not perfect. Most obviously, their size is limited and even if they were otherwise perfect the RMS error in the mean of an m-member ensemble would

converge to $\sqrt{(1+m^{-1})}$ times the mean-square analyzed anomaly (Leith, 1974). It means that if the model and the 10-member ensemble investigated here were otherwise perfect the RMS (MS) for a large sample of forecasts would converge with increasing lead time to $\sqrt{1.1}$ (1.1) times the mean-square analyzed anomaly.

Based on the above discussion, a condition for the optimal smoothing by a finite size ensemble can be defined in the long time limit. The problem is to find a condition that can be used at shorter forecast lead times. To derive a condition of this type it is useful to rescale Eq.3 by the mean-square analyzed anomaly

$$NMS = 1 + (NFA - 2 \times AC \times \sqrt{NFA}) \equiv 1 - SKILL. \tag{4}$$

Here, the *normalized MS (NMS)*, the *normalized mean-square forecast anomaly (NFA)*, and the *anomaly correlation (AC)* are, respectively,

$$NMS = \frac{<(f-a)^2>}{<(a-c)^2>}, \tag{5}$$

$$NFA = \frac{<(f-c)^2>}{<(a-c)^2>}, \tag{6}$$

$$AC = \frac{<(f-c)(a-c)>}{\sqrt{<(f-c)^2><(a-c)^2>}}. \tag{7}$$

We note that this expression for AC, which can also be found in the textbook of Wilks (1995), is equal to the correlation between the forecast and analyzed anomalies only if the mean of the forecast and analyzed anomalies ($<(f-c)>$ and $<(a-c)>$) are negligible.

MS is smaller for a forecast f than for the climatology based forecast c, or in other words, NMS can be smaller than one, if and only if $SKILL > 0$. This inequality is satisfied for a given value of AC whenever $0 < \sqrt{NFA} < 2 \times AC$, and then the largest possible reduction in NMS is achieved when

$$\sqrt{NFA} = AC \tag{8}$$

and in that case

$$SKILL = AC^2. \tag{9}$$

Eq. 9 was first derived by Murphy and Epstein (1989), but in a somewhat different context. They assumed, as is the case for a good single NWP model forecast, that NFA is equal to one, hence, except for the initial time, $SKILL$ is *smaller* than AC^2. Therefore, they argued, the square of the anomaly correlation should be considered as a measure of *potential* rather than *actual* skill. Here we make use of the fact that by taking an ensemble mean, NFA gradually decreases with increasing forecast lead time. This means that ensemble forecasting provides a way of turning the *potential skill* provided by AC^2 of the ensemble mean to *actual skill*. It must also be emphasized that ensemble averaging can also increase AC, thus increasing the potential skill itself. In other words, a change in the EFS improves the SKILL if it leads to the retention of more skillful features with more realistic amplitude and/or to the more efficient filtering of unpredictable details. The optimality condition we were searching for is given by Eq. 8.

The above arguments can be extended to give an estimate of the RMS reduction that can be attributed to the reduced forecast anomalies in the ensemble mean compared to the control forecast. Let us assume that AC is equal for the ensemble and the control forecasts, its value is ac and the value of NFA for the control forecast is nfa. The NMS is smaller for the ensemble mean than for the control forecast if and only if $SKILL - (2 \times ac \times \sqrt{nfa} - nfa) > 0$. This condition is satisfied if and only if $\sqrt{NFA}$ is between (i) $2 \times ac - \sqrt{nfa}$ and (ii) $\sqrt{nfa}$, while the ensemble mean has the largest possible advantage over the control if Eq. 8 is satisfied. In most cases, including the one when $nfa = 1$, condition (i) provides the lower and condition (ii) the upper bound, but when AC is high and nfa is smaller than one (the numerical model is unrealistically diffusive) the regular order of the bounds may be reversed. Since the smallest meaningful value of $\sqrt{NFA}$ is zero, that should be used as lower bound whenever condition (i) gives a negative negative number. If the anomaly correlation (ac) is high (close to 1) there is only a narrow range of reduced forecast error variance (NFA) that can improve the rms error (NMS) and a strong smoothing can only degrade the forecast quality (RMS). When the anomaly correlation (ac) is low, however, the similarly low value of condition (i) provides

a wide range of forecast variance reduction (NFA) that can lower the rms error (NMS).

3.2 Relative RMS error

A *Relative RMS*

$$RRMS = \frac{RMS_h}{RMS_{T62}} \times 100 \tag{10}$$

is introduced here as the percentage of the RMS in the initially high resolution runs (RMS_h) compared to that in the T62 run (RMS_{T62}). For small RMS values $RRMS$ is a more sensitive measure than the difference between the RMS of the two forecasts ($RMS_{T62} - RMS_h$), thus it can be more efficiently used to analyze the RMS reduction. Throughout this section the relative RMS is used in figures to demonstrate changes in the forecast quality due to increased model resolution. Values smaller (larger) than 100 indicate forecast improvement (degradation) due to increased horizontal resolution.

3.3 RMS and AC for the control forecasts

NH RRMS (Fig. 2): All initially high resolution control runs have lower RMS error. Furthermore, the use of reduced model resolution after 1 day has a clear negative effect on the quality of the control forecasts for the day 2-4 forecast range. Reducing the resolution after 3 days, however, has an opposite effect in the 4-7 day time range by improving the scores. The benefit from increased resolution is largest (7.3% error reduction) at 2-day lead time, after which the gain in forecast quality gradually diminishes.

SH RRMS (not shown): The initial impact of increased resolution on the control forecast is positive but less significant than for the NH region. For the day 2-4 range the T126 and the D3 runs are superior to the D1 run again, but beyond day-4 the T126 has the largest rms error.

NH AC (Fig. 3, D1 and D3 are not shown): The evaluation of AC suggests similar conclusions to those drawn based on RMS statistics. The AC is higher for the runs that were started

at resolution T126 than for the low resolution control at all forecast lead times. The D3 run produced consistently higher scores again than the T126 run for the day 4-7 forecast range.
SH AC (not shown): The T126 run is superior to the others only up to 4 days.

3.4 RMS and AC for the ensemble means

NH RRMS (Fig. 4): The improvement in the skill of the ensemble mean forecasts due to increased resolution is more substantial than that for the control forecasts. Interestingly, the error reduction is largest (10.1%) at the shortest verified lead time (1-day), in contrast to the case of the control forecasts, for which the largest error reduction was observed at day-2. Another important difference is that the T126 forecast remains superior to the D3 run for the day 4-7 forecast range, too.
SH RRMS (not shown): The ensemble mean shows a behavior similar to that of the control forecast: the maximum benefit of increased resolution is at day-2 lead time; the T126 and the D3 runs are superior to the D1 run for the day-2-4 forecast range; and the T62 forecast outperforms the T126 run after 4 days.
NH AC (Fig. 3), D1 and D3 are not shown because they are almost indistinguishable from those for the T126 forecasts:. The initially high resolution runs performed better than the T62 reference ensemble in terms of AC, too. Note that the advantage of the higher resolution runs at day 7 for the ensemble mean is about 12 forecast hours compared to only about 2 hours for the control. This means that the increased resolution had a more positive impact on the potential skill of the ensemble mean than on that of the control forecast.
SH AC (not shown): The initially high resolution mean forecasts have an advantage in terms of AC for the first four days. At day-5 the AC for the different runs is identical, while beyond that time the T62 run has the highest AC values.

3.5 Comparison of the ensemble means and controls

NH NMS and RMS (Fig. 5, RMS is not shown): In the T62 forecasts the RMS is 6.5% larger for the mean forecast than for the control at day-1 and only after 4 days does it become smaller. In contrast, for the initially high resolution runs the ensemble mean RMS is larger only at day-1, and only by a small amount (less than 1.5%), while beyond that time the mean has a gradually increasing advantage over the control.

The NMS for the T126 (T62) control reaches the error level of the climatological forecast at 9-day (8.5-day) lead time. The advantage of the ensemble mean over the control at this time is 32.7% (29.9%), what is equivalent to a 18% (16.3%) RMS reduction. At 15-day lead time the NMS is 0.903 for the ensemble mean, an indication that the ensemble mean provides a forecast typically better than that based on climatology even at this extended forecast range. This is consistent with earlier results shown in Zhu et al. (1996) and Toth et al. (1998).

SH NMS and RMS (not shown): The RMS (MS) is lower for the means than for the corresponding control forecasts at all lead times. The error level of the forecast based on climatology is reached at around 6-day lead time by both the T62 and the T126 control forecasts. The advantage of the means at this time is 30.4% (28.1%) for the T126 (T62) ensemble, which is very similar to that observed for the NH region. The controls and the means converge to their asymptotes by day-12 indicating that there are no predictable features in the SH region beyond that time.

NH AC (Fig. 3): While the AC indicates a lower potential skill for the T62 mean than for its control during the first 3 forecast days, the AC is always higher for the initially high resolution mean forecasts than for their controls.

SH AC (not shown): The AC scores indicate that the mean forecasts have higher potential skill than their controls.

Summary. The increased horizontal resolution has *more positive effect on the mean than on the control forecasts* in the NH region. This is in part due to the fact that for the first 3 days the forecast quality of the T62 mean is lower than that of the control. Truncating

the fields after 3 days of model integration has no benefit in the case of the mean forecasts, which is in contrast to the behavior observed for the control forecasts. This indicates that the unpredictable small scale features are efficiently filtered by the ensemble average and an additional non-selective spectral filtering of the meteorological fields degrades the forecast quality.

The skills of the ensemble mean and the corresponding control forecasts are more similar in the SH region and there is no obvious advantage of integrating the ensembles at increased horizontal resolution beyond 4 days.

3.6 Time evolution of NFA

NFA is first evaluated for the control forecasts (not shown). As can be expected from a good NWP model, NFA remains near one for all controls during the entire 15-day forecast range. More precisely, in the NH region there is a slight (less than 0.5%) initial decay of NFA during the first day, but beyond that time the mean-square distance between the forecasts and climatology is practically perfect. In the SH region the initial decay is somewhat more pronounced (still less than 2.5%) and NFA shows a slow growing trend for the longer lead times, which explains why the long term limit of MS is larger than 2.

Fig. 6 presents the time evolution of NFA for the T126 and the T62 ensemble mean forecasts in the NH region. The mean-square forecast anomaly in the NH region is considerably larger for the T62 than for the T126 run at all forecast lead times. The fact that the T62 model produces higher NFA than the T126 model for the ensemble mean is also indicated by the relatively high NFA for the D1 and D3 runs: after the resolution is reduced there is a well distinguishable sudden increase in NFA (not shown).

The optimal value of NFA, AC^2 (Eq. 8) , is also shown in Fig. 6 for both ensembles. NFA for the T126 run is optimal at day-1, after which it gradually drifts away from its optimum. This is not surprising if we recall that the long term limit of NFA would be 0.1 in a 10-member ensemble. At day-15 the asymptotic values are not established yet, but the

difference between NFA (0.17) and its possible lowest value (0.1) is small.

The behavior of the T62 ensemble mean is drastically different. The most striking feature is the curious result that NFA is the largest for the T62 ensemble mean among all forecasts including the controls. This can happen only if the 24-hour ensemble mean contained new structures that were non-existent in the other runs. Nonlinearity must play an important role since linearly evolving perturbations would result in identical ensemble mean and corresponding control forecasts. This behavior will be discussed in more detail in section 5.

Apparently, the RMS could be improved by rescaling the forecast anomalies $(f - c)$ such that NFA would become optimal. The RMS reduction that could be achieved by such a rescaling would be 1.2% (2.8%, and 3.3%) at 3-day (7-day and 10-day)forecast lead time for the T62 ensemble, while the same number for the T126 ensemble is less than 0.1% (0.7% and 1.1%). The RMS values for the T126 ensemble mean are nearly as good as they can be for a 10-member ensemble. This indicates that structures retained in the ensemble mean usually *have realistic magnitude and significant improvements in the mean forecasts can be expected only if AC can be further improved.* This also means that for the high resolution ensemble mean *the actual skill is almost equal to the potential skill* given by AC^2.

4 Forecast bias and forecast error variance

4.1 Decomposition of the mean-square error

In this part of the study, all variables are decomposed into two parts. A mean forecast (analyzed) flow, $\bar{f}$ ($\bar{a}$), is defined by time averaging the forecasts (analyses) for each lead time separately, over the 30-day sample period. Eddy quantities, f' (a') are then computed by taking the deviation of the forecast (analyzed) values from the forecast (analyzed) mean. The local mean-square error can also be decomposed by partitioning the meteorological fields into mean and eddy components

$$MS(\lambda, \phi) = \overline{(f-a)^2} = \overline{(\bar{f}-\bar{a})^2} + \overline{(f'-a')^2} = \overline{f-a}^2 + \overline{(f-a)'^2}. \tag{11}$$

The square-root of the first term on the rhs. of Eq. 11 (the error in the predicted mean) is conventionally called *bias*, while the second term (the error in the predicted eddy component) is called the *forecast error variance*. In Figures 7 and 8 the two error components for the T126 control run are shown at 1-day and 4-day lead times. The localized patterns associated with large bias have negligible contribution to the total error along the storm track regions. Meanwhile, a close relationship between storm tracks and the dominant patterns of short-term forecast error variance is already well established at day-1 and it becomes even more evident at day-4. This means that decomposing MS into bias and variance components provides a good way to define a forecast error component that is dominantly determined by errors in the prediction of high-frequency transients at short lead times.

The global MS can also be decomposed into bias and error variance terms by taking the average of Eq. 11 over the grid points

$$MS = < (f-a)^2 > = < \overline{(f-a)}^2 > + < (f-a)'^2 > . \tag{12}$$

In what follows the relative importance of the increased resolution in reducing the bias versus the error variance component of MS is investigated by plotting figures based both on Eqs. 11 and 12. When reduction in the local bias (forecast error variance) is concerned the difference between the square of the bias (forecast error variance) terms for the low and the high resolution runs are mapped. Positive (negative) values on these maps mark locations of reduced (increased) error terms. In case of the space-time-averaged bias (forecast error variance) a relative bias (forecast error variance) is defined by the ratio of the square of the the bias (forecast error variance) terms for the initially high resolution and the T62 runs. Like in the case of the relative RMS shown in earlier figures, the ratio is expressed in percentage and the bias (forecast error variance) is reduced whenever the relative bias (forecast error variance) is smaller than 100.

4.2 Time evolution of the forecast bias component

NH forecast bias, control (Figs. 9 and 10): The strongly localized patterns of large forecast bias are over or near to land at day-1. In particular, at the location where the largest error reduction occurs in the region of the Tian-Shan mountains (in central Asia), there is an extremely large gradient in surface height. An independent study (H. M. van den Dool and S. Saha 1999, personal communication) concluded that this is the typical location of the largest short term forecast bias in the NCEP global forecast of the 500 hPa height. The same authors, using the technique of Empirical Orthogonal Teleconnections (van den Dool et al., 2000), found that the error at that location is also in a close relationship with the short term coherent large scale Northern Hemisphere error patterns in the NCEP (but not in the ECMWF) global model. By day-4 the strongly localized patterns of improvement disappear and a mixture of modest amplitude improvements and degradations is left behind.

Since the magnitude of the local bias reductions is large the increased resolution has a dramatic positive influence on the control forecast. The spatially averaged error reduction (Fig. 10) is the largest (35%) at the shortest verified lead time and the spectral truncation has an immediate negative impact. Truncating the forecast after day-1 has a clear negative impact, while reducing the resolution after day-3 is less damaging but still clearly negative.

SH forecast bias, control (not shown): The improvements in the *SH* region are smaller than in the NH region. The largest error reduction (16%) is at day-3. The bias reduction rapidly decreases beyond day-4 and completely diminishes by day-6.

NH forecast bias, ensemble mean (Fig. 11 and 12:) The dominant improvements are concentrated in the same region as for the control forecast at 1-day lead time. Similar, but smaller magnitude improvement patterns (not shown) can be observed by comparing the T126 and the D1 (D3) runs at day-2 (day-4). The impact of increased resolution is more dramatic than in the case of the control forecast: the error reduction is 56% at day-1 and day-2 and the error reduction is still significant (29%) at day-7.

SH forecast bias, ensemble mean (not shown): The error reduction shows a similar trend to

that observed for the control forecast in the SH region. The only difference is that the error reduction for the short forecast lead times is larger for the ensemble mean than for the control.

4.3 Time evolution of the forecast error variance

NH error variance, control (Figs. 13 and 14): Large errors in the prediction of transient eddies along the mid-latitude storm tracks were significantly reduced by the increased resolution. There are also significant error reductions in regions where the T126 control forecast produced no significant error variance but where it was burdened by large forecast bias. The most obvious example for this is the large-bias area in the Tian-Shan mountains. It means that in the T62 model not only the forecast means but also the forecast transients are in error in the above regions. While truncating the forecasts after day-1 has negative forecast effects in the day-2-4 forecast range, the D3 run is superior to the untruncated T126 run at and beyond the day-4 forecast lead time. The largest error reduction, 15%, is achieved at day-3 by the T126 forecast.

SH error variance, control (not shown): The initially high resolution forecasts are less efficient in reducing the variance component of the error than in the NH region. Also, the D1 and D3 runs clearly outperform the T126 run in that (1) the error for the T126 beyond day-4 is even larger than that for the T62 run; (2) the largest error reduction, 6%, is achieved by the D1 run at day 2 lead time; and (3) the truncation always has an immediate positive impact on the error variance component.

NH error variance, ensemble mean (Figs. 15 and 16): The main areas of error reduction are in the storm track regions and over Europe. The largest spatially averaged error reduction, 12% achieved by the T126 forecast at day-3, is somewhat smaller for the ensemble mean than for the control forecast. On the other hand, for the longer than 5-day forecasts, the error reduction is larger for the mean than for the control and the T126 mean outperforms the truncated D3 mean.

SH error variance, ensemble means (not shown): The positive effect of increased model

resolution is relatively modest. The largest error reduction, which is only 4.9% at day-3, was realized by the T126 run. The D1 run performs slightly worse in the day 2-4 range, while the D3 run is somewhat better in the day 4-7 range than the T126 run.

5 Discussion

Transient eddies. The verification statistics indicate that increased horizontal resolution enhances the model performance, partly as expected, through better handling of the transient eddies. Since the predictability limit is shorter than three days for a large group of the transient features, a spectral truncation of the forecasts at those lead times can improve the control forecast performance both in RMS and AC terms. Ensemble averaging, on the other hand, removes a large part of the unpredictable details from the forecasts. This ensures that the T126 ensemble mean, in contrast to the T126 control, remains superior to its truncated counterpart even for the medium and the extended forecast ranges in the $NHregion$. Though most of the improvements in the prediction of the high frequency transients can be realized by integrating the forecasts at high resolution only out to day-3 or even only to day-1, a reduction in the resolution at these times has a clear negative impact on the mean forecasts.

Over the *SH region* the forecast error variance in both the control and the mean forecast is reduced even after truncating the resolution at as early as 1-day lead time. The most plausible explanation for this is that the less adequate data coverage results in a relatively poorer analysis of the smaller scales, leading to an earlier loss of predictability, and an elevated level of forecast error variance for these scales. The rather modest reduction of error variance found in the ensemble mean and the much shorter time limit for skillful prediction in the SH region seem to confirm this explanation. We note also that because of the poorer quality verifying data sets (analyses) the verification results themselves are less reliable in the SH than in the NH region.

Forecast bias. The less anticipated result is that the significantly larger bias in the T62 resolution prediction of the 500 hPa height field plays an important role in explaining the

difference between the quality of the different resolution control and mean forecasts at short lead times. Those components of the predicted phase space trajectories which are associated with grid-points in the problematic regions rapidly drift toward the artificial climate of the T62 model. This drift leads to an inevitable increase of the local and the global MS (RMS) error statistics. Moreover, because $(\bar{f}-c)^2$ is much larger than $(\bar{a}-c)^2$ in the large-bias regions, NFA is also larger in the presence of bias. This is because

$$NFA = \frac{<(f-c)^2>}{<(a-c)^2>} = \frac{<(\bar{f}-c)^2> + <f'>}{<(\bar{a}-c)^2> + <a'>} \tag{13}$$

and in our case, as in any good analysis-forecast system $<f'>$ and $<a'>$ are nearly equal. This latter requirement is satisfied by both the T62 and the T126 versions of the NCEP MRF except for a slight deficit in $<f'>$ for the first few days. In the case of the control forecasts this deficit is sufficient to compensate the short-term impact of the bias on the NFA, but the growing trend of the bias in the SH region eventually leads to the increase of NFA, reported in section 3.6. The high value of NFA for the T62 ensemble mean in the NH region (which is even larger than one at 1-day lead time) can also be explained by the presence of bias. For this forecast the difference between $(\bar{f}-c)^2$ and $(\bar{a}-c)^2$ is significantly larger than for any other mean or control forecast verified in this paper.

It is not clear whether our results point to a specific problem with the T62 version of the NCEP MRF model or they are more general indicating that a realistic flow cannot be maintained in all geographic areas in a T62 resolution NWP model. Our results suggest that doubling the horizontal resolution to T126 can eliminate much of the systematic errors, though slight improvements from further resolution increases can still be expected.

Model errors and ensembles The overall improved performance of the higher resolution ensemble can be explained as a combined effect of better predicting the transient eddies and significantly reducing the forecast bias.

Our challenge is to find an explanation for the anomalously large forecast bias in the T62 ensemble mean. It is obvious that the time evolution of the perturbed trajectories in the T62 ensemble is highly nonlinear; otherwise the mean and the control forecasts were

identical, which would be reflected in identical forecast scores as it is almost the case for the T126 resolution forecasts. The presence of perturbations obviously plays a role in the curious behavior of the T62 ensemble mean. We should keep in mind that the initial perturbations in the T62 and the initially T126 resolution ensembles are generated exactly the same way except for the use of different resolution models. That is, the anomalous behavior of the T62 ensemble mean somehow must be related to the way the perturbations and the model bias interact.

We know that the bred perturbations (rescaled difference between pairs of short range forecasts) consist of perturbation patterns that amplify fastest in a cycle of 1-day forecast differences. It has been argued previously (Szunyogh et al. 1997, Toth and Kalnay, 1997) that the bred vectors consist of unstable structures dominantly associated with baroclinic instabilities of the atmosphere, as represented by the numerical models. The results of the present study suggest that in addition to the above structures, another type of rapidly amplifying structures, related to the drift of the model from the analyzed state to a field that the imperfect T62 model is able to maintain, are also present in the bred perturbations. There is a crucial difference, however, between the behavior of the model drift induced structures and those related to atmospheric instabilities. The former is the artifact of the use of an imperfect model, while the latter is the result of properly modeled real world processes. The inclusion of realistic unstable structures, as ensemble initial conditions, leads to a nonlinear error reducing process in the ensemble mean. The inclusion of structures that induce model drift, however, leads to aggrevated errors in the ensemble mean.

6 Conclusions

We conclude with the following observations:

- The increased horizontal resolution enhances the performance of the ensemble mean forecasts. The rms error for the Northern Hemisphere mid-latitude mean forecast is reduced for the entire 15-day forecast range, while the anomaly correlation is increased for the first

11 days. In the Southern Hemisphere mid-latitudes the same error statistics are improved for the first four days of model integration.

• The balance between anomaly correlation and forecast variance is more optimal in the T126 than in the T62 ensemble mean. In other words, using a higher resolution model the actual skill is closer to the potential skill defined by AC^2.

• The two main meteorological aspects of the resolution induced error reductions are the maintenance of a more realistic time-mean flow and the better prediction of high frequency transients along the mid-latitude storm tracks.

• The effect of increased horizontal resolution is more positive on the ensemble mean than on the control forecast. The maximum rms error reduction for the ensemble mean (control) is 10.1% (7.3%) in the Northern Hemisphere mid-latitudes. This improvement is found at 1-day (2-day) forecast lead time, but at day-7 the error reduction is still 5.2% (1.9%). At day-7 the advantage of the high resolution ensemble mean (control) in terms of anomaly correlation is 12 (2) hours. In the Southern Hemisphere the rms error reduction for the mean (control) is 3.6% (3.5%) at day-2.

• It is evident that the adequate model resolution is most crucial during the first few days of model integration. While the ensemble clearly benefits from maintaining high resolution beyond the first three days both in terms of reduced bias and error variance, the error variance in the control forecast is actually reduced when the resolution is truncated at day-3.

While some of the quantitative results presented in this study may strongly depend on the sample period chosen, the indications are clear that using adequate model resolution in ensemble forecasting is important. The results shown here, along with probabilistic verification scores to be presented in a separate study, demonstrate that the use of a higher resolution NCEP MRF leads to improved ensemble forecasting. Partly based on the results presented here, a new operational ensemble configuration was implemented at 1200 UTC 27 June 2000 (1200 UTC 20 December 2000) at NCEP. Since this implementation ten perturbed forecasts are made both at 0000 and 1200 UTC and all perturbed forecasts are integrated at a hori-

zontal resolution T126 up to 60-hour (84-hour), after which, in order to save computer time, they are truncated to T62 resolution.

Acknowledgments

We would like to thank the staff of EMC, and in particular Drs. Stephen Lord and Hua-Lu Pan for their support. Yannick Tremolet and Joe Sela of NCEP provided valuable help with setting up the replica of the global EFS on the new Class-VIII computer of NCEP, while Yuejian Zhu helped with forecast verification. Glen White and Jun Du of NCEP provided helpful comments on an earlier version of this manuscript. The authors are also grateful to Anders Persson of ECMWF for enjoyable discussions on forecast verification.

References

Buizza, R., J. Barkmeijer, T. N. Palmer, and D. S. Richardson, 1999: Current status and future developments of the ECMWF Ensemble Prediction System. *Met. Apps.*, **6,** 1-14.

Buizza, R., T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingswoth, A. Simmons, and N. Weidi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Q. J. Roy. Meteorol. Soc.,* **124,** 1935-1960.

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.,* **123,** 2181-2196.

Iyengar, G., Z. Toth, E. Kalnay, and J. S. Woolen, 1996: Are the bred vectors representative of analysis errors? *Preprints of the 11th AMS Conference on Numerical Weather Prediction*, August 19-23, 1996, Norfolk, VA, 64-65.

Kadar, B., I. Szunyogh, and D. Devenyi, 1998: On the origin of model errors. Part II. Effects

of the spatial discretization for Hamiltonian systems. *Idojaras,* **102,** 71-107.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437-471.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409-418.

Machenauer, B., 1991: Spectral methods. In *Proc. ECMWF Seminars on Numerical methods in atmospheric models,* 9-13 Sept. 1991, ECMWF, Shinfield Park, Reading RG2 9AX, United Kingdom, 3-85. Molteni, F., R. Buizza, T.N. Palmer and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.,* **125,** 3241-3270.

Murphy, A. H. and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572-581.

Richardson, D. S., 2000: Skill and economic value of the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.,* **126,** 649-668.

Simmons, A. J., R. Mureau, and T. Petroliagis, 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Q. J. Roy. Meteorol. Soc.,* **121,** 1739-1771.

Szunyogh, I., Z. Toth, R. E. Morss, S. J. Majumdar, B. J. Etherton and C. H. Bishop, 2000: The effect of targeted dropsonde observations during the 1999 Winter Storm Reconnaissance program. *Mon. Wea. Rev.,* **128,** 3520-3537.

Szunyogh, I., E. Kalnay, and Z. Toth, 1997: A comparison of Lyapunov and optimal vectors in a low-resolution GCM. Tellus, **49A,** 200-227.

Talagrand, O., Vautard, R., and Strauss, B., 1999: Evaluation of probabilistic prediction systems. In *Proc. ECMWF Workshop on Predictability,* 20-22 Oct. 1997, ECMWF, Shinfield Park, Reading RG2 9AX, United Kingdom, 1-26.

Toth, Z., Y. Zhu, and R. Wobus, 2000: On the economic value of ensemble based weather forecasts. *Bull. Amer. Meteorol. Soc.,* under review.

Toth, Z., Y. Zhu, T. Marchok, S. Tracton, and E. Kalnay, 1998: Verification of the NCEP global ensemble forecasts. Preprints, 12th Conference on Numerical Weather Prediction,

11-16 January 1998, Phoenix, Arizona, 286-289.

Toth, Z. and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297-3319.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NCEP: The generation of perturbations. *Bull. Amer. Meteorol. Soc.,* **74,** 2317-2330.

Tracton, S. and E. Kalnay, 1993: Ensemble forecasting at NCEP: Operational implementation. *Wea. Forecasting,* **8,** 379-398.

White, G. H., 1999: Systematic errors in NCEP operational global analysis/forecast system. Preprints, *13th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 94-95.

Wilks, D. S., 1995: *Statistical methods in the atmospheric sciences.* Academic Press, San Diego, pp. 467.

Wilks, D. S., and Hamill, T. M., 1995: Potential economic value of ensemble-based surface weather forecasts. *Mon. Wea. Rev.,* **123,** 3564-3575.

Zhu, Y, G. lyengar, Z. Toth, M. S. Tracton, and T Marchok, 1996: Objective evaluation ofthe NCEP global ensemble forecasting system. Preprints, 15th AMS Conference on Weather Analysis and Forecasting, 19-23 August 1996, Norfolk, Virginia, p. J79-J82.

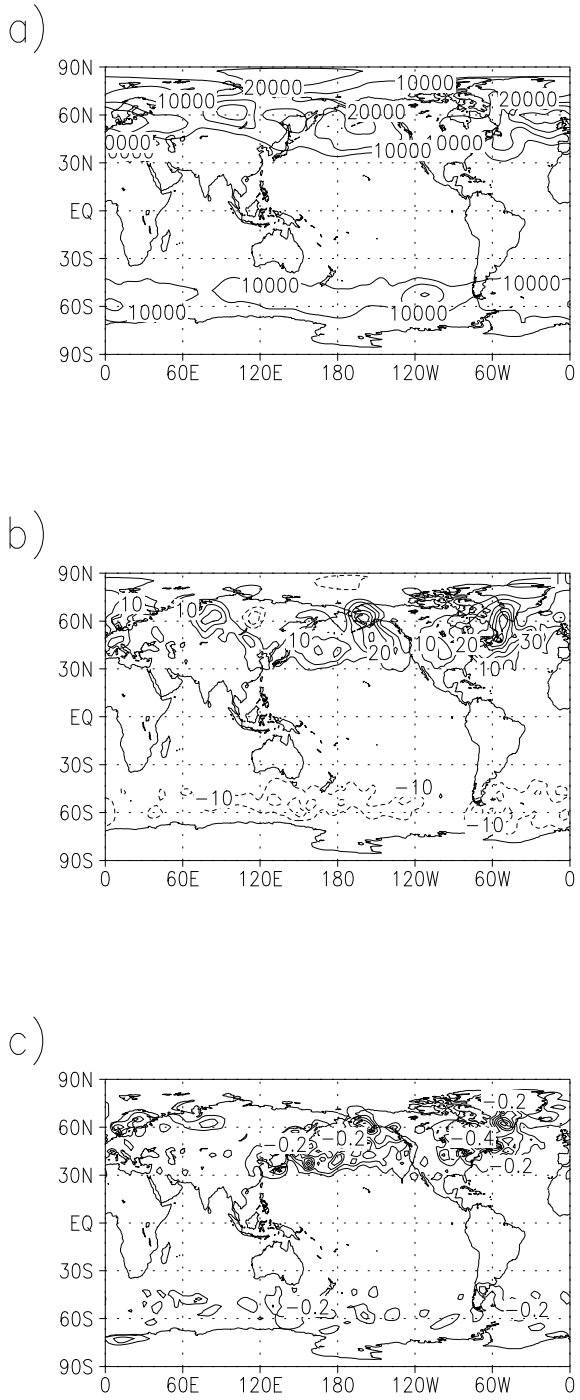

Figure 1: Eddy statistics. (a) Geopotential height variance at the 500 hPa pressure level. Contour interval is 10000 gpm^2. (b) Meridional temperature flux at the 700 hPa pressure level. Contour interval is 10 Km/s. (c) Vertical temperature flux at the 700 hPa pressure level. Contour interval is 0.2 KPa/s.

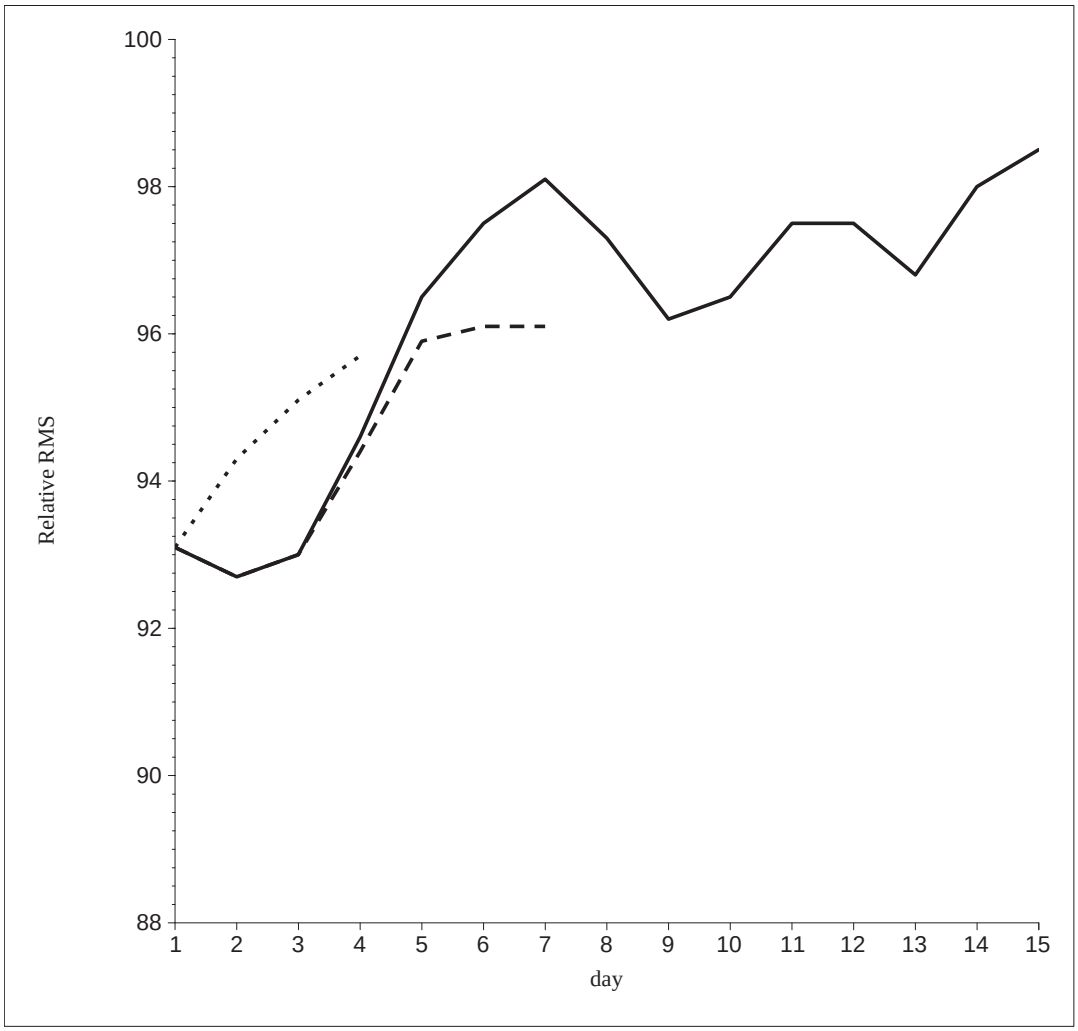

Figure 2: Relative RMS for the T126 (solid line), the D3 (long dashes) and the D1 (short dashes) control forecasts.

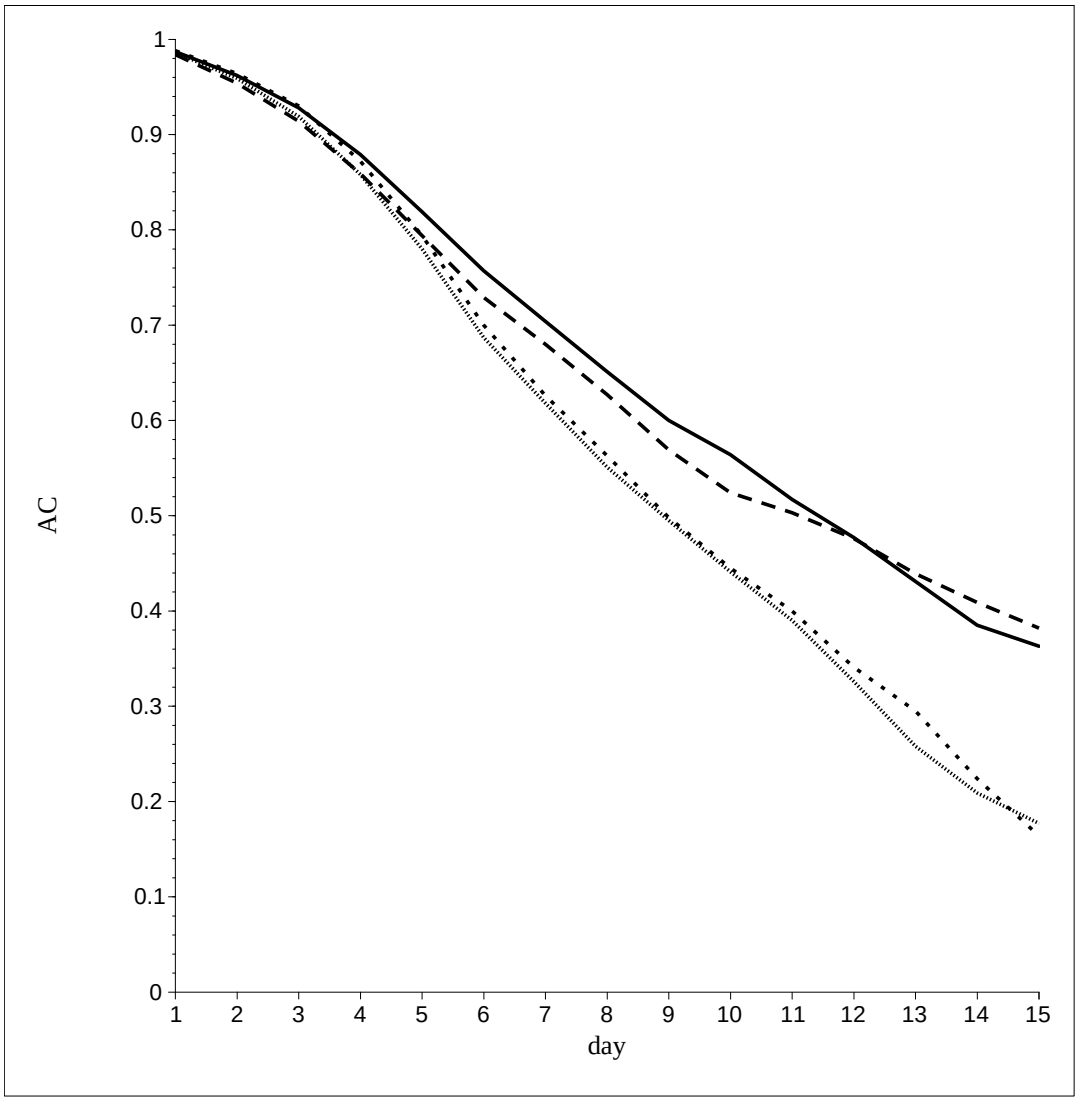

Figure 3: Anomaly correlation for the T126 control (short dashes) and mean (solid line) and the T62 control (dotted line) and mean (long dashes) forecasts.

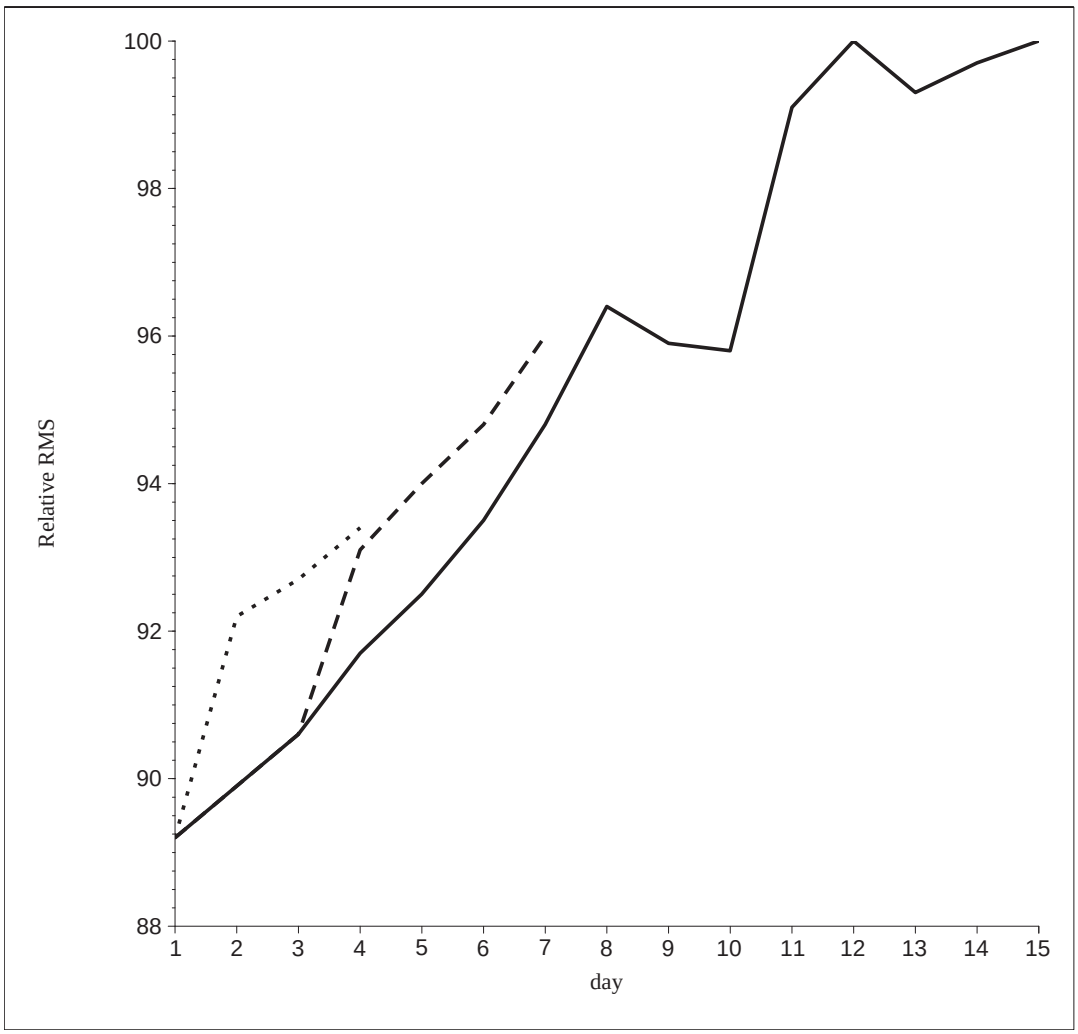

Figure 4: Same as Fig. 2 but for the ensemble means.

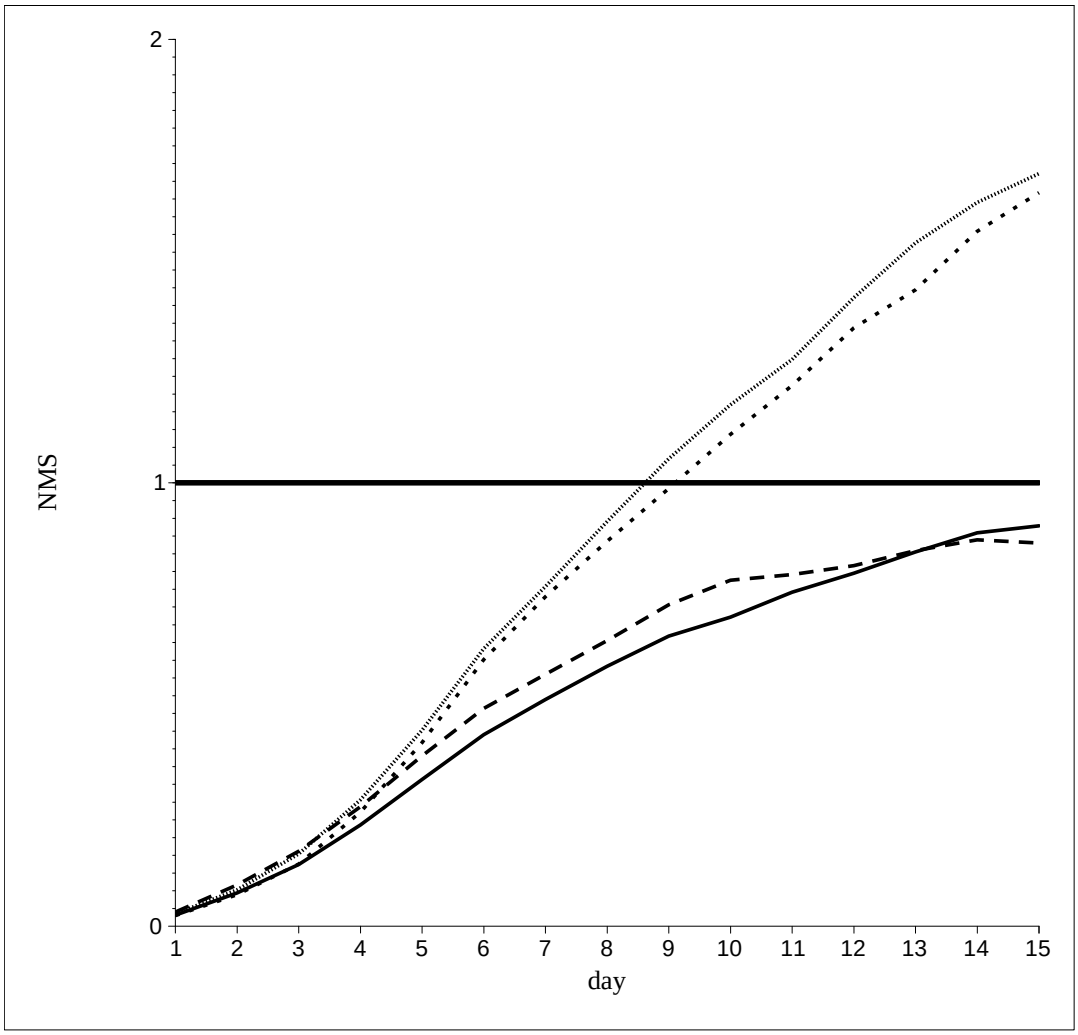

Figure 5: Normalized mean-square error (NMS) for the T126 control (short dashes) and mean (solid line) and the T62 control (dotted line) and mean (long dashes) forecasts. The thick solid line shows the normalized mean-square error for the climatology.

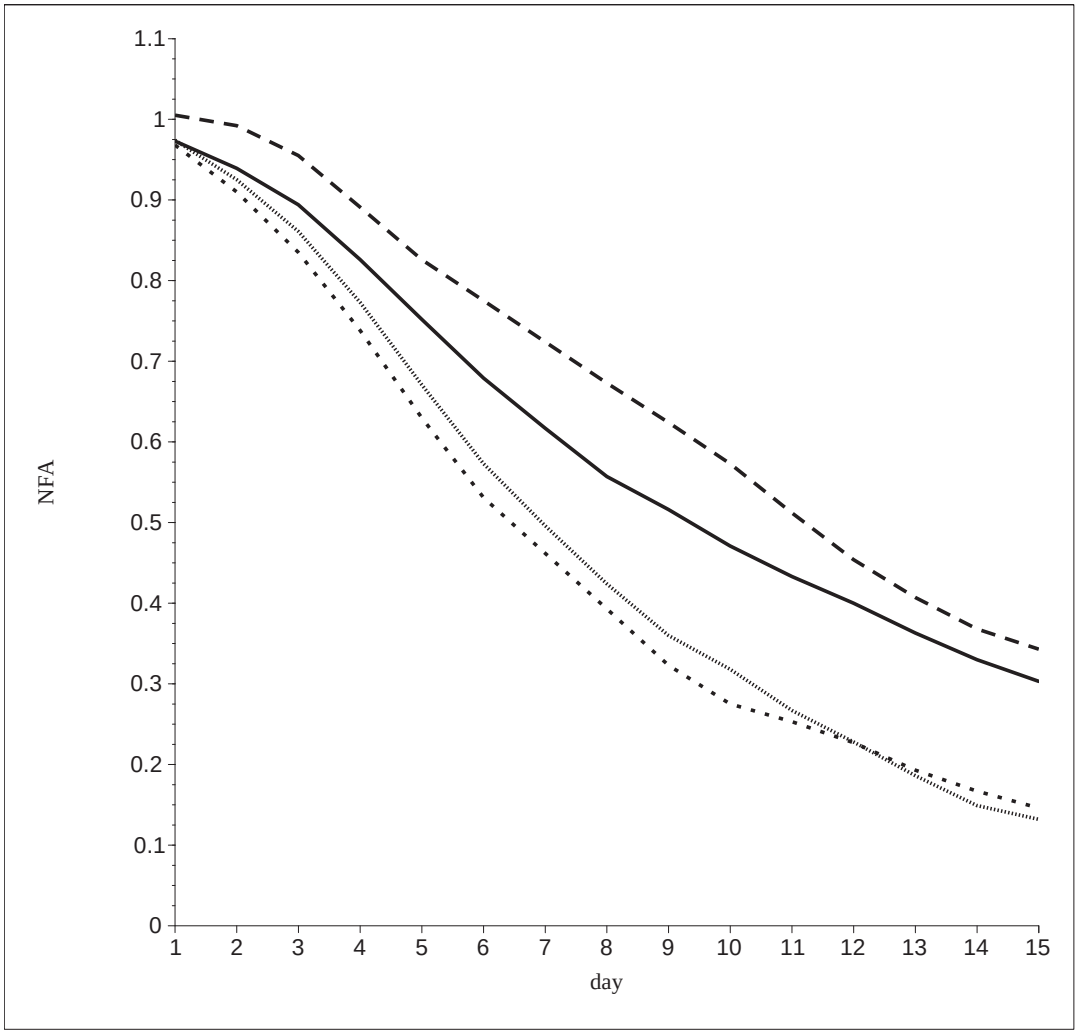

Figure 6: Normalized mean square distance (NFA) between the T126 ensemble mean and climatology (solid line). Long dashes show the same but for the T62 mean. The hypothetical optimal value (AC^2) is shown by dotted line (short dashes) for the T126 (T62) ensemble.

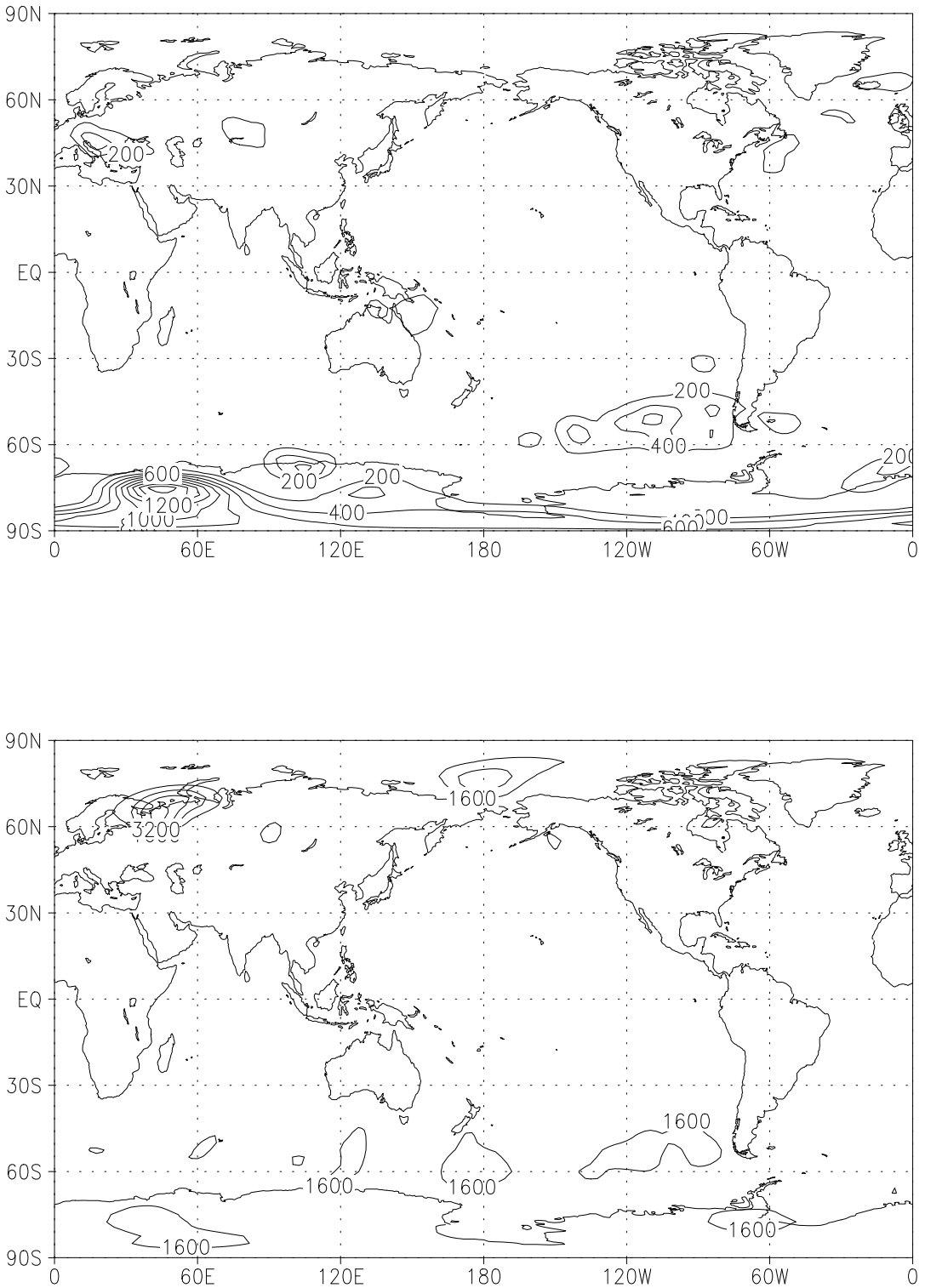

Figure 7: Square of the local bias, $\overline{f-a}^2$, for the T126 control run at day-1 (upper panel, contour interval is 200 gpm^2) and at day-4 (lower panel, contour interval is 1600 gpm^2) forecast lead times.

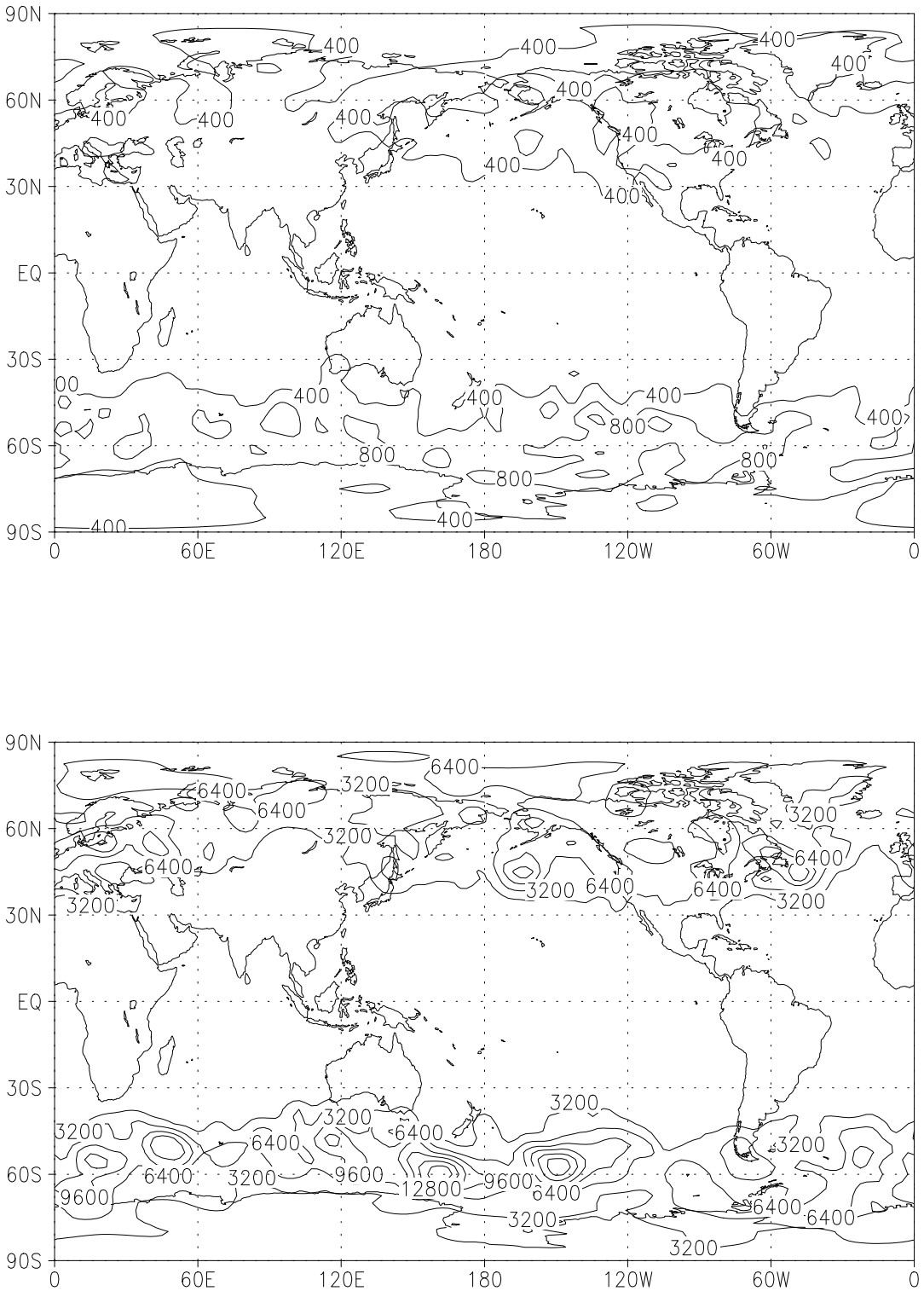

Figure 8: Forecast error variance, $\overline{(f-a)'^2}$, for the T126 control run at day-1 (upper panel, contour interval is 400 gpm^2) and at day-4 (lower panel, contour interval is 3200 gpm^2) forecast lead time.

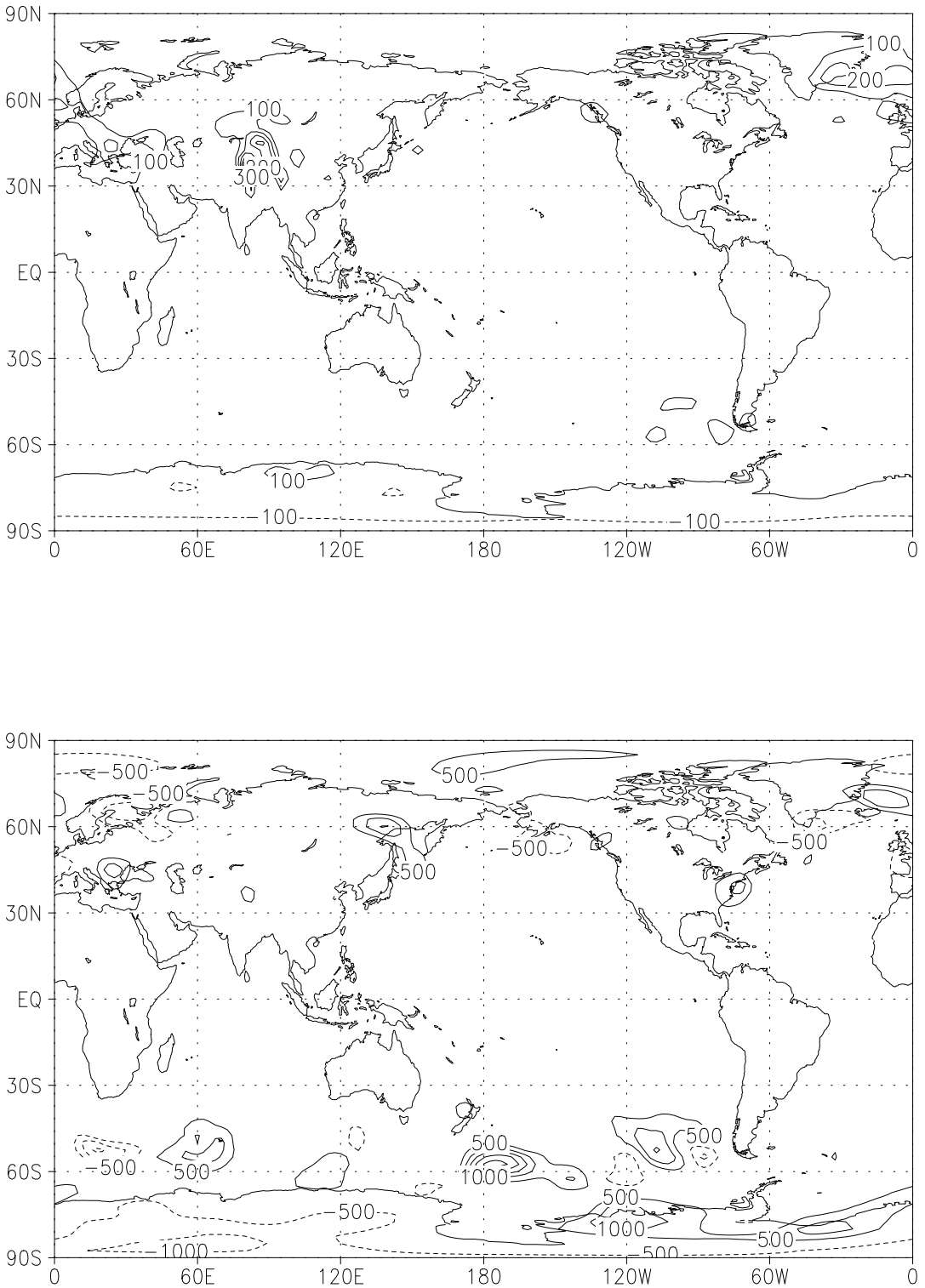

Figure 9: The difference between the square of the local bias for the T62 and the T126 control forecasts at day-1 (upper panel, contour interval is 100 gpm^2) and at day-4 (lower panel, contour interval is 500 gpm^2). Positive values mark regions where the bias is smaller for the T126 than for the T62 control.

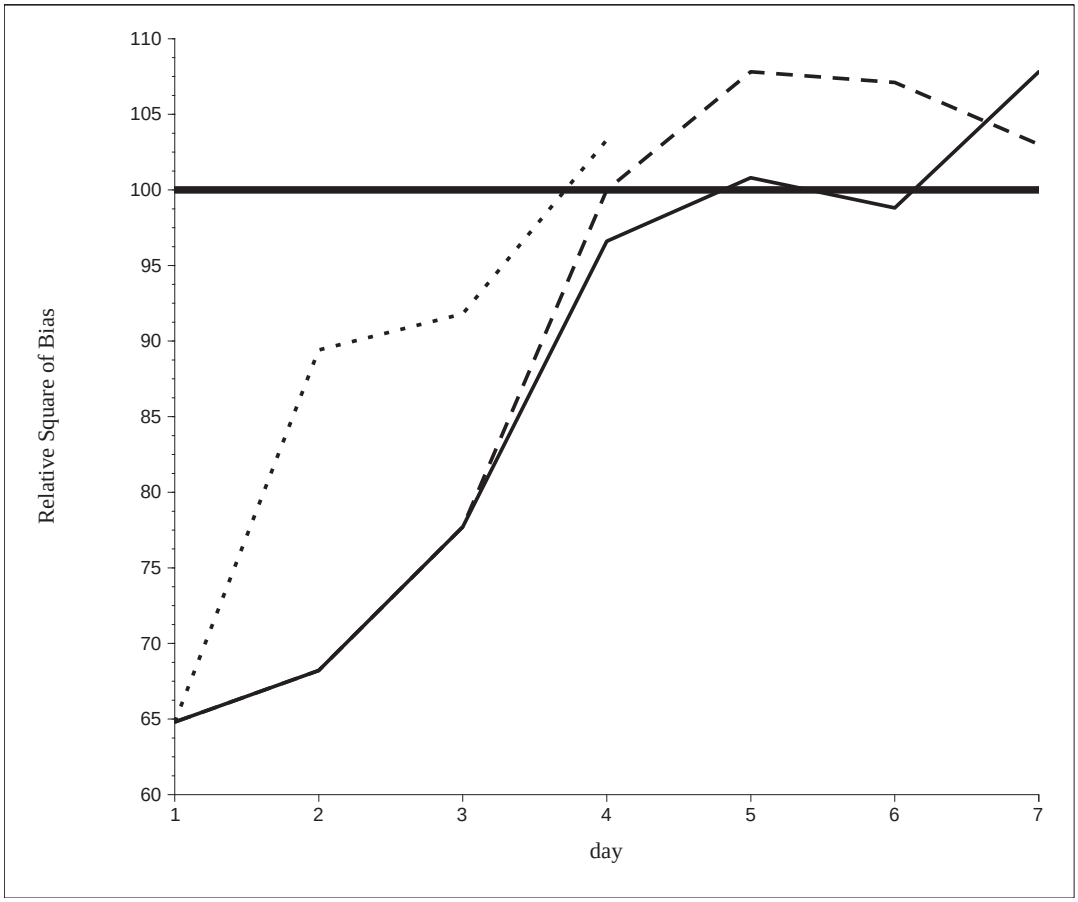

Figure 10: Relative square of the bias for the T126 (solid line), the D3 (long dashes) and the D1 (short dashes) control forecasts. The thick solid line shows the relative square of bias for the T62 control forecast.

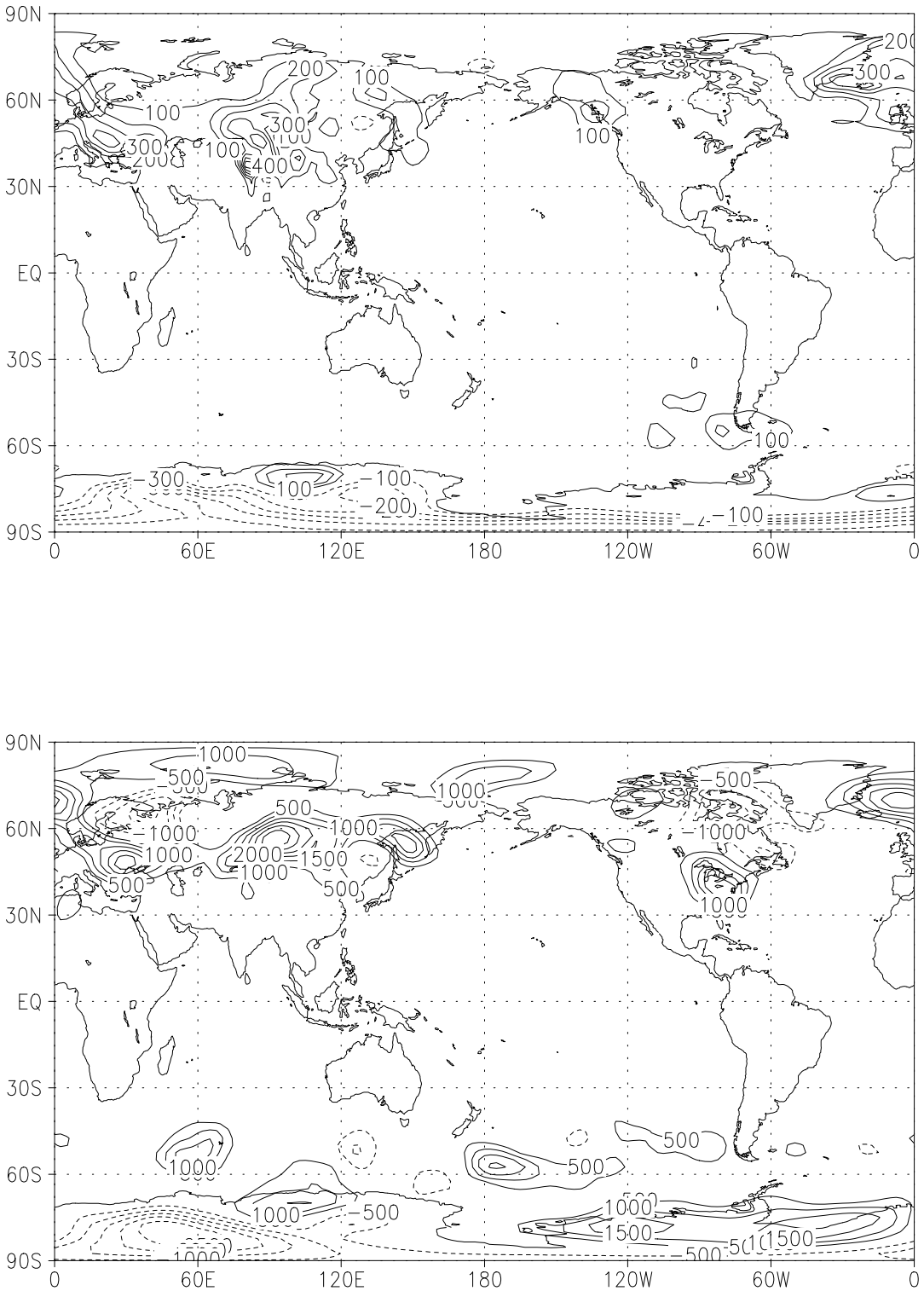

Figure 11: The difference between the square of the local bias for the T62 and the T126 mean forecasts at day-1 (upper panel, contour interval is 100 gpm^2) and at day-4 (lower panel, contour interval is $500 gpm^2$). Positive values mark regions where the bias is smaller for the T126 than for the T62 control.

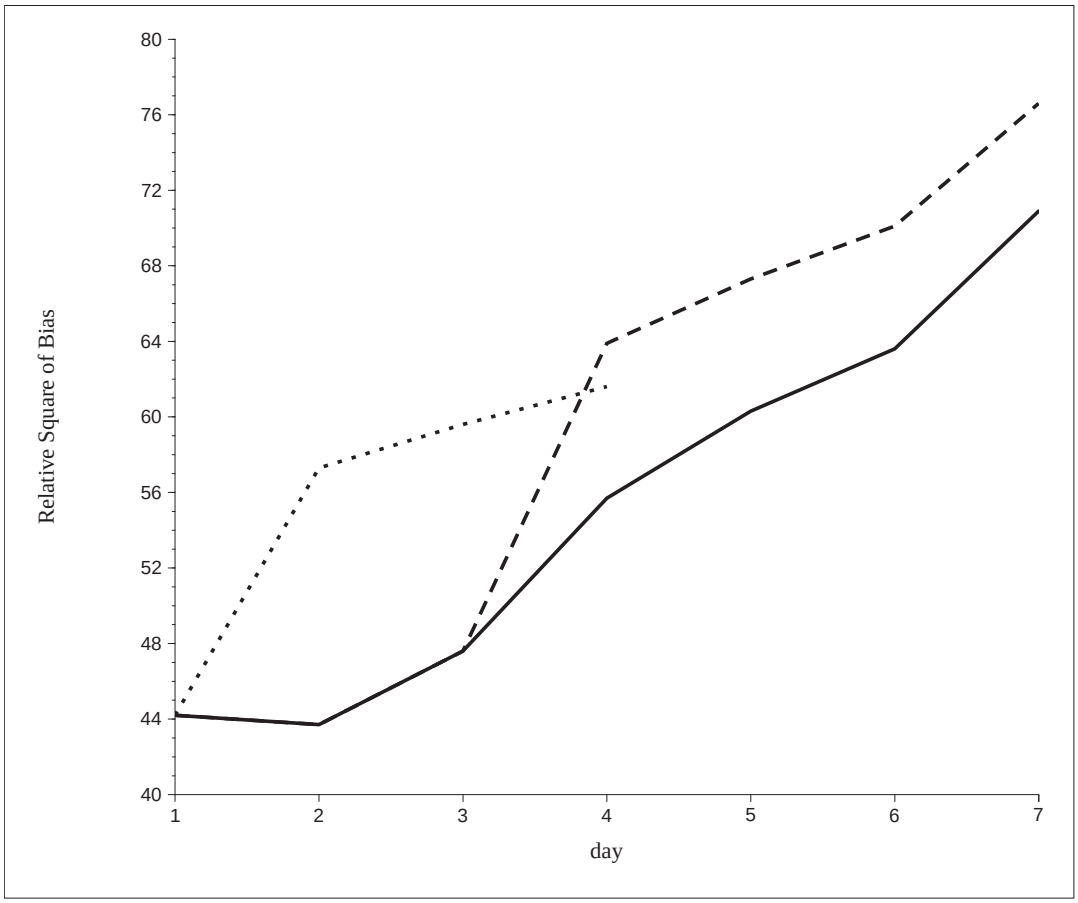

Figure 12: Relative square of the bias for the T126 (solid line), the D3 (long dashes) and the D1 (short dashes) mean forecasts.

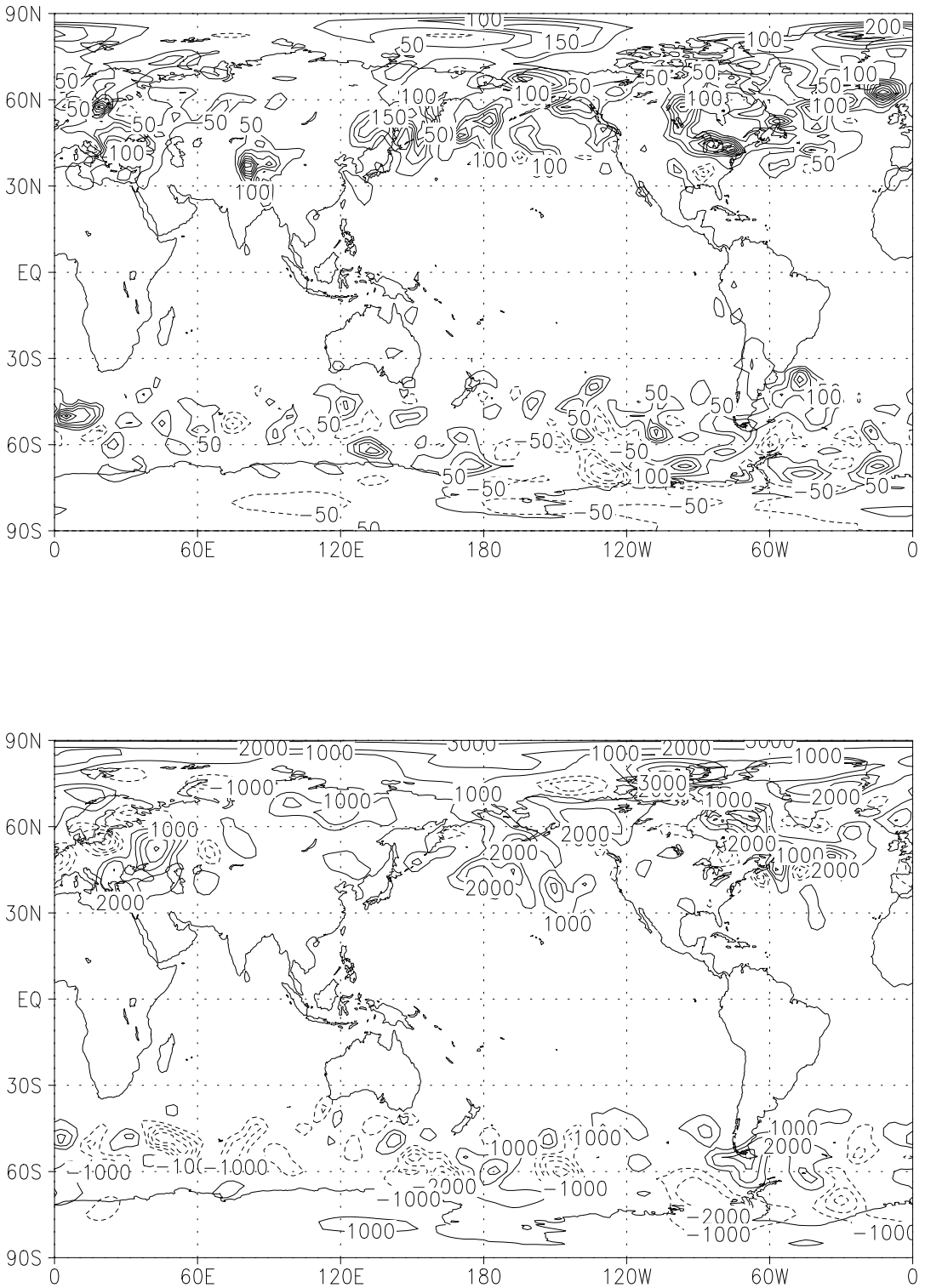

Figure 13: The difference between the forecast error variance for the T62 and the T126 control forecasts at day-1 (upper panel, contour interval is 50 gpm^2) and at day-4 (lower panel, contour interval is 1000 gpm^2). Positive values mark regions where the forecast error variance is smaller for the T126 than for the T62 control.

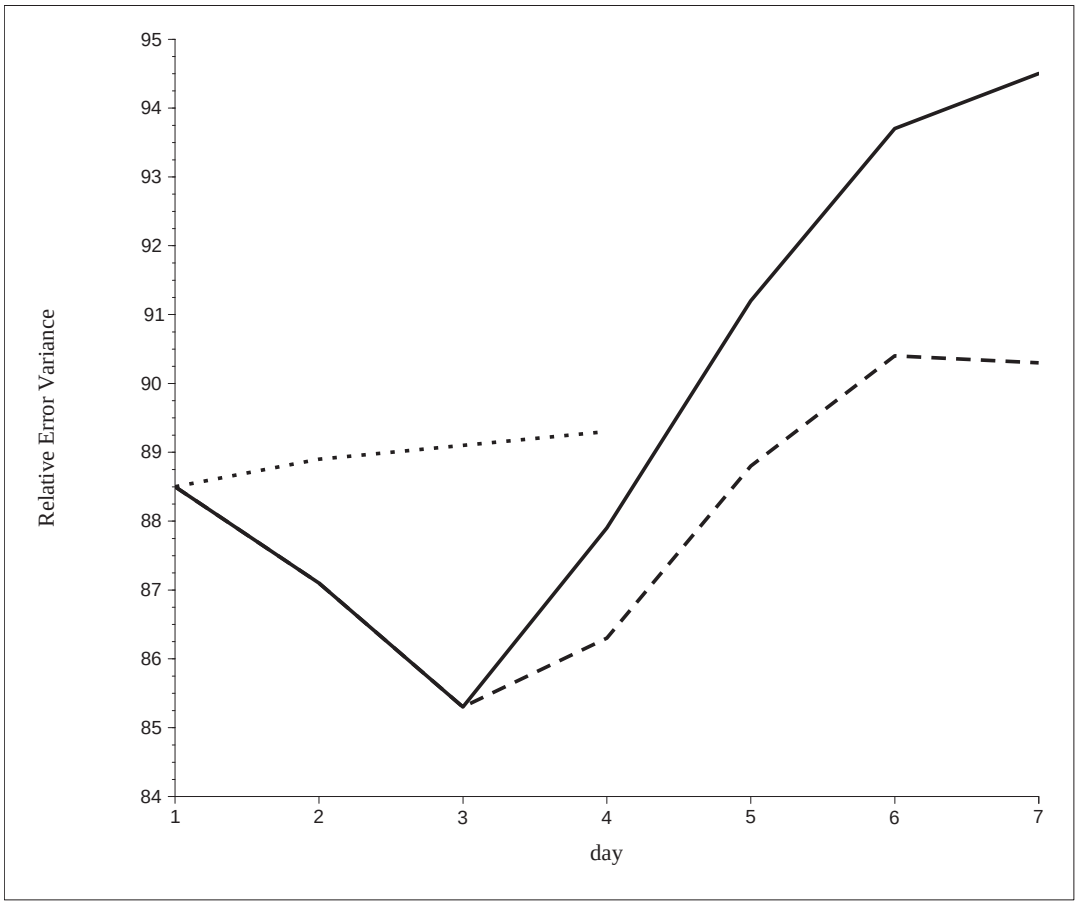

Figure 14: Relative forecast error variance for the T126 (solid line), the D3 (long dashes) and the D1 (short dashes) control forecasts.

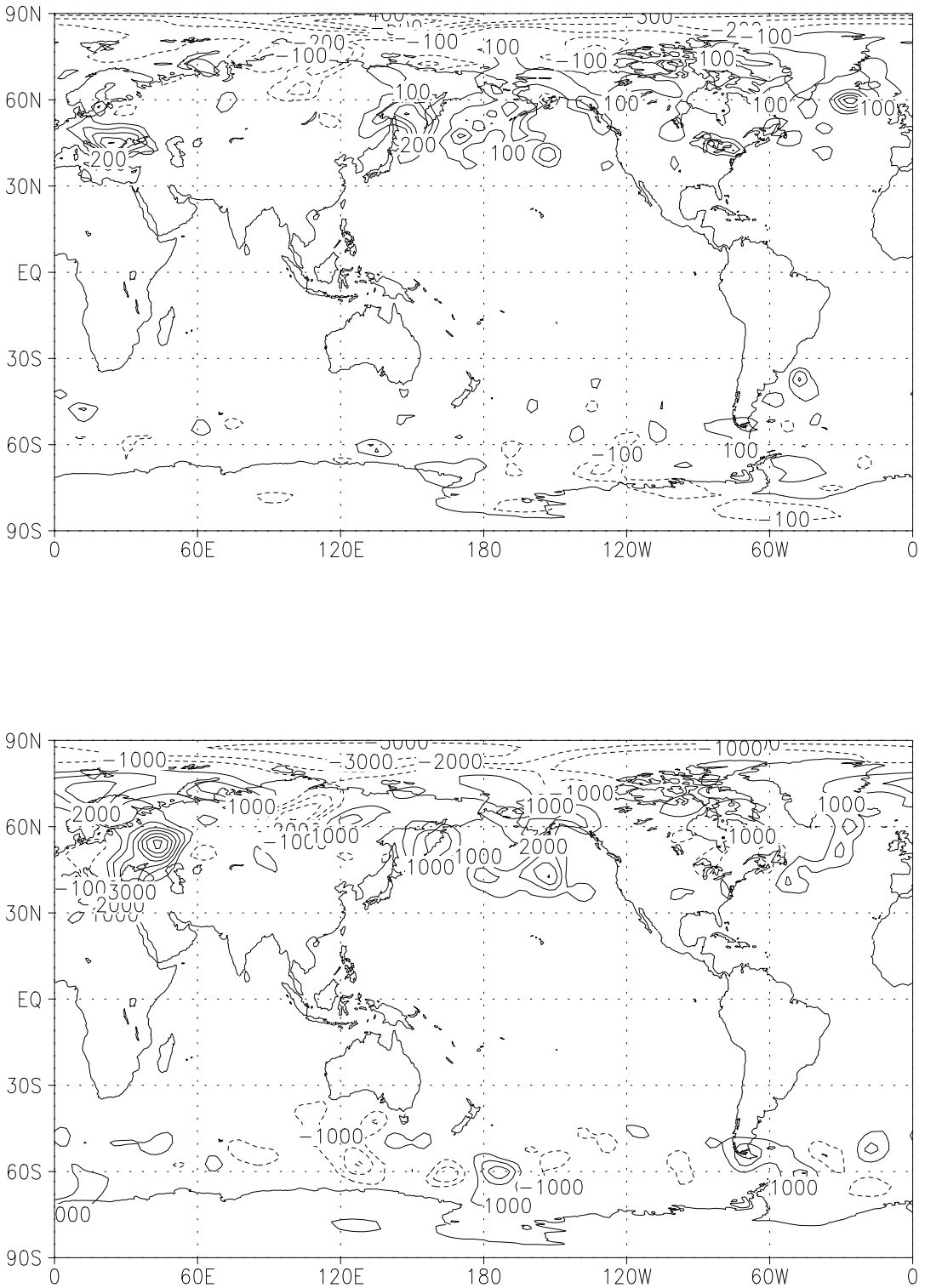

Figure 15: The difference between the forecast error variance for the T62 and the T126 mean forecasts at day-1 (upper panel, contour interval is 100 gpm^2) and at day-4 (lower panel, contour interval is 1000 gpm^2). Positive values mark regions where the forecast error variance is smaller for the T126 than for the T62 mean.

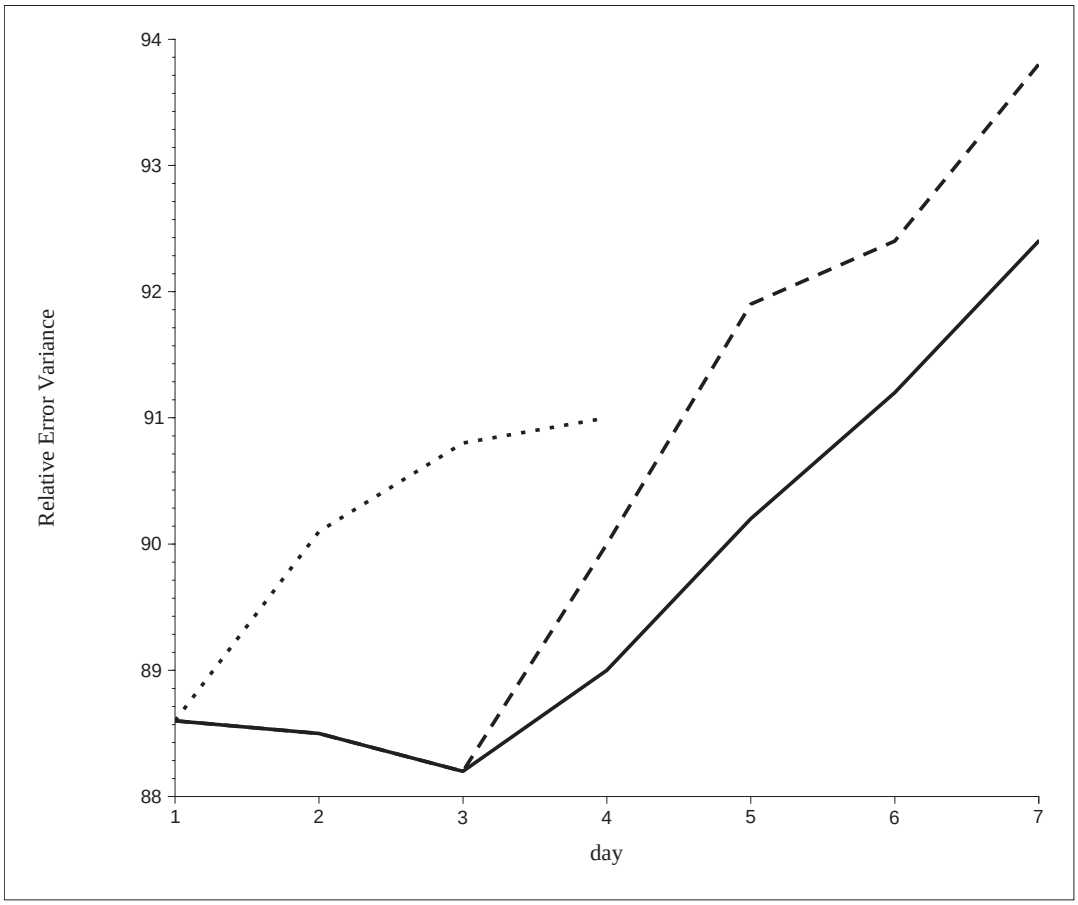

Figure 16: Relative forecast error variance for the T126 (solid line), the D3 (long dashes) and the D1 (short dashes) mean forecasts.