Using Filtering to Mitigate Stochastic Model Errors' Effect on Ensemble Covariance. Part I: Evaluation of a Prototype Filtering Scheme.

Justin G. McLay[1]

NRC/Naval Research Laboratory, Monterey, California

Jonathan E. Martin

University of Wisconsin, Madison, Wisconsin

(Submitted to Monthly Weather Review, November 29, 2004)

[1] Corresponding Author: Justin G. McLay, Naval Research Laboratory, 7 Grace Hopper Ave., Stop 2, Monterey, CA, 93943-5502. E-mail: mclay@nrlmry.navy.mil.

ABSTRACT

This paper is the first in a two-part series that investigates a method of ensemble post-processing designed to mitigate the effect of stochastic model errors on ensemble covariance. The method performs a series of filtering experiments with the operational ensemble members, obtaining a set of forecast states which is less corrupted by errors of stochastic origin. It then uses some number of the filtered states to complement or supplant the operational members, forming a so-called hybrid ensemble.

In this paper the method is introduced and a prototype filtering scheme is established for use in the method's evaluation. The efficacy of the filtering scheme is gauged through composite and ensemble-by-ensemble comparisons between the root-mean-square error characteristics of the set of filtered states and those of the operational ensemble. These comparisons are based upon a year's worth of global ensemble data, and suggest that the filtering scheme can in fact consistently produce a set of states which are generally less corrupted by stochastic errors than the operational members are. Some underlying properties of the prototype filtering scheme are also highlighted. This is accomplished by investigating the filtered states' performance from the perspectives of probabilistic modelling and so-called anomaly projection.

1. Introduction

In numerical weather prediction the initial state can only be fully described in terms of some probability density function (PDF). Also, arbitrarily small errors in the initial state may exhibit large growth. These two observations effectively eliminate any chance at successful deterministic weather forecasts except on those rare occasions wherein the forecast evolution is relatively insensitive to initial state uncertainty. In view of this reality, weather forecasts should be probabilistic in nature. Ensemble prediction originated as a means to estimate the forecast PDF given some initial PDF associated with the uncertainty in the specification of the initial state (Leith 1974, Molteni et al. 1996, Toth and Kalnay 1997). In practice, the method involves performing a series of model integrations from some finite set of initial conditions obtained through perturbation of the best estimate of the initial state. The set of forecast states provided by the integrations defines the estimate of the forecast PDF. Probability forecasts for arbitrary weather events can be derived from the estimate of the forecast PDF.

The effectiveness of an ensemble prediction system is dependent upon two factors: how well the initial PDF is defined and sampled, and how well the numerical model simulates the dynamics and physics of the real atmosphere. The greater proportion of early efforts into ensemble prediction was devoted to defining and sampling the initial PDF. A large body of recent work has suggested, however, that model errors can have considerable detrimental impact on the skill of ensemble-derived probability forecasts (Colucci and Baumhefner 1998, Buizza et al. 1999, Harrison et al. 1999, Palmer et al. 1999, Stensrud et al. 1999, Evans et al. 2000, Palmer 2000, Orrell et al. 2001, Palmer 2001, Wandishin et al. 2001, Alhamed et al. 2002, Bright and Mullen 2002, Mylne et al. 2002, Barkmeijer et al. 2003, Orrell 2003). Thus, it has become clear that if the utility of ensemble-derived probability forecasts is to be fully realized, then the characteristics and sources of model errors must be identified, and means of mitigating or eliminating these errors must be devised and implemented.

For elaboration on the model error problem consider two different forecast trajectories, one evolved from a given initial state using a numerical model and one evolved from the same initial state using real atmospheric dynamics and physics (Fig. 1). The difference at some time t between the state on the numerical model's forecast trajectory and the state on the real atmospheric forecast trajectory represents the error introduced by the model as of time t (Fig. 2). This error may have two components, one of systematic origin and one of stochastic origin. The error component associated with systematic processes has the same direction and magnitude for each ensemble member (Fig. 3). This uniformity results in the mean of the ensemble distribution being displaced from that of the real atmospheric distribution. The ensemble distribution is said to be "biased" in this circumstance. The error component associated with stochastic processes does not necessarily have the same direction and magnitude for each ensemble member (Fig. 4). The effect of this error component is to corrupt the variance and covariance of the ensemble distribution. For illustration, consider that the covariance of two random variables X and Y is expressed as

$$Cov\,(X,Y) \;=\; E\,[XY] \;-\; E\,[X]\;E\,[Y]\;, \tag{1}$$

where $E\,[]$ is the expectation operator. If stochastically-induced errors a and b are added to X and Y, respectively, then the covariance of the resulting corrupted versions of X and Y (referred to as X^* and Y^*) is

$$Cov\,(X^*,Y^*) \quad = \quad Cov\,(X,Y) \;+\; E\,[Xb] \;+\; E\,[Ya] \;+\; E\,[ab]\;, \tag{2}$$

where it is assumed that the expected value of both a and b is zero. Thus, stochastically-induced errors will alter the covariance of random variables through the last three terms on the right-hand side of Eq. 2.

Methods of dealing with model systematic and stochastic errors in ensemble prediction can be classified as to whether they operate during the course of the numerical integrations or after the integrations are completed. Methods of the former class include those of Buizza et al. (1999) and Bright and Mullen (2002), which incorporate additional terms in the numerical model to mitigate the effect of stochastic errors in the model's parameterized coupling between unresolved- and resolved-scale flow. Also in the former class is the "forcing singular vector" method of Barkmeijer et al. (2003), which assesses the effect of systematic perturbations to model-variable physical tendencies.

Methods that compensate for model errors after the numerical integrations are completed are referred to as "post-processing" methods. One of these, termed the calibration method, adjusts ensemble-derived event probabilities in accordance with some distribution (parametric or nonparametric) that describes the historical likelihood that the verifying event will assume a certain location in the hierarchy of ensemble-depicted events (Hamill and Colucci 1998, Eckel and Walters 1998, Hamill 2001). Calibration is able to mitigate the effects on ensemble-derived probabilities of both model systematic errors and incorrect ensemble variance (which is partly the product of model stochastic errors). Another post-processing method is the formation of a multi-model ensemble (Harrison et al. 1999, Stensrud et al. 1999, Evans et al. 2000, Wandishin et al. 2001, Mylne et al. 2002, Alhamed et al. 2002). This involves the combination of ensemble members from two or more different ensemble prediction systems, on the basis that the members from one system may not be subject to the same model systematic and stochastic error effects as the members from the other system(s). What might be termed a derivative form of the multi-model ensemble method is the perturbed-model ensemble method (Houtekamer et al. 1996, Stensrud 2000). In this approach, the ensemble members are all generated using the same model, but each member is generated using different combinations of that model's available physics packages, convective schemes, etc. Another, emerging post-processing method is that of ensemble member dressing (Roulston and Smith 2003, Wang and Bishop 2004). This approach involves forming a statistical ensemble for each member of the dynamical ensemble (i.e. the ensemble obtained by integrations of a numerical model) by sampling the space around the given member in accordance with the member's historical error statistics. The statistical ensembles (described as "dressing") are then combined with the dynamical ensemble to form a so-called hybrid ensemble, from which probabilistic forecasts are henceforth derived. The dressing method provides some compensation for the effect of stochastic model errors, and its development has established the feasibility and economy of using statistical, or otherwise non-dynamical, samples in ensemble prediction.

It is interesting to note that none of the existing methods for mitigating model error effects employ filtering. As the objective of filtering is to retrieve the underlying good information from a corrupted signal, it is not unreasonable to think it might be useful in some capacity for dealing with forecast trajectories that are contaminated by model errors. It is also interesting that all but one of the existing methods are intended to improve the mean and/or variance of the ensemble distribution, but not necessarily its covariance. The ensemble dressing method of Wang and Bishop is the exception, as it strives for improvement in ensemble covariance by ensuring that the statistically generated members' covariances are identical to the dynamical members' seasonally averaged error covariances. Since the problems with ensemble distributions' statistical moments have proven resilient to individual solutions, the existence of but one method for improving ensemble covariance is impetus for investigating alternative such methods. Given these observations, a natural question is whether a post-processing technique that is based upon filtering can be designed for the purpose of mitigating the effect of stochastic model errors on ensemble covariance. The current paper and McLay and Martin (2005) take the opportunity to set forth and assess one technique that is being explored in response to this question. Along with introducing the new technique, these papers also have the broader objective of opening a new line of inquiry into the challenging problem of stochastic model errors.

The organization of the current paper is as follows. A synopsis of the post-processing technique is provided in Section 2. In Section 3, a prototype filtering scheme is defined to facilitate investigation of the technique. Section 4 presents a comprehensive assessment of the prototype filtering scheme's ability to produce a set of states with reduced-amplitude stochastic errors, based upon a large sample of ensembles. In Section 5, some underlying properties of the prototype filtering scheme are revealed through examination of the question of whether certain filtered states systematically perform better than others. Conclusions are presented in Section 6.

2. Description of the Method

Expressed mathematically, the basic objective is to find filtered versions of the corrupted random variables X^* and Y^* (referred to as ${X_f}^*$ and ${Y_f}^*$, the subscript f identifying a filtered variable) such that

$$|\, Cov\left({X_f}^*, {Y_f}^*\right) \;-\; Cov\left(X, Y\right) \,| \;<\; |\, Cov\left(X^*, Y^*\right) \;-\; Cov\left(X, Y\right) \,| \;, \tag{3}$$

where the vertical bars denote an absolute value. To further examine this relation, recognize that $Cov\left({X_f}^*, {Y_f}^*\right)$ can be represented in analogy to Eq. 2 as

$$Cov\left({X_f}^*, {Y_f}^*\right) \quad = \quad Cov\left(X_f, Y_f\right) \;+\; E\left[X_f b_f\right] \;+\; E\left[Y_f a_f\right] \;+\; E\left[a_f b_f\right] \;, \tag{4}$$

where it is assumed that the filtering scheme is distributive and that $E\left[a_f\right] = E\left[b_f\right] = 0$. For ease of presentation the three terms in Eqs. 2 and 4 that involve stochastic errors collectively will be referred to as e and e_f, respectively, such that

$$Cov\left(X^*, Y^*\right) \quad = \quad Cov\left(X, Y\right) \;+\; e \;, \qquad \text{and} \tag{5}$$

$$Cov\left({X_f}^*, {Y_f}^*\right) \quad = \quad Cov\left(X_f, Y_f\right) \;+\; e_f \;. \tag{6}$$

Substituting the righthand sides of Eqs. 5 and 6 into Eq. 3 for $Cov\left(X^*, Y^*\right)$ and $Cov\left({X_f}^*, {Y_f}^*\right)$, respectively, then taking advantage of a rule of absolute value and assuming that $E\left[X_f\right] = E\left[X\right]$ and $E\left[Y_f\right] = E\left[Y\right]$ it is found that Eq. 3 can be expressed as

$$|\, E\,[X_f Y_f] \;-\; E\,[XY]\,| \;<\; |e| \;-\; |e_f|\;. \qquad (7)$$

Eq. 7 allows for some insight into the conditions that must be met if filtering is to be used to improve ensemble covariance. Two observations follow upon its inspection:

1. The righthand side of Eq. 7 must be greater than zero. This means that the filtering must produce a distribution of states with stochastic errors that are generally of reduced amplitude, so that $|e_f| < |e|$.

2. The lefthand side of Eq. 7 ideally would be zero. This circumstance is most easily realized if the filtering causes no reduction in amplitude of the true[1] states X and Y. However, because both sides of Eq. 7 involve differences it is apparent that some filtering of the true states is permissible, provided that some reduction in amplitude of the stochastic errors is achieved simultaneously. This means that the filtering does not have to exactly differentiate between stochastic errors and true states in order to provide for improved covariance. What the filtering does have to do is reach an effective compromise between the elimination of errors and the smoothing of true flow states.

It must be borne in mind that the points made regarding the second observation above rest in part on the aforementioned assumption that $E\,[X_f] = E\,[X]$ and $E\,[Y_f] = E\,[Y]$. This assumption essentially represents another condition that must be satisfied if the filtering is to prove beneficial. In the case that $E\,[a_f] = E\,[b_f] = 0$ (a not unreasonable scenario) then this condition will be met simply if $E\,[{X_f}^*] = E\,[X^*]$ and $E\,[{Y_f}^*] = E\,[Y^*]$.

A filtering scheme that satisfies the conditions discussed above can facilitate an improvement in ensemble covariance. The main impediment to using filtering for this purpose is the reduction in amplitude of true flow states that is attendant with the filtering process. This reduction in amplitude will operate to increase the lefthand side of Eq.

[1] In the context of numerical weather forecasts, a "true" state is one that would be realized if an initial state is evolved with the dynamics of the real atmosphere.

7 and thereby degrade estimates of covariance. Similarly, it will serve to diminish estimates of variance. In fact, even if filtering can engender an improvement in covariance, this improvement may not translate to an improved multi-dimensional probabilistic forecast because of the detriment of diminished variance. The hypothesis that underlies the present analysis is that one might be able to mitigate the problem of the filtering of true states without eliminating the possibility of improved covariance by using some number of the filtered states in concert with some number of the unfiltered states in a so-called "hybrid" ensemble distribution. Adoption of this hybrid ensemble approach means that a two-part post-processing methodology will be explored in the present analysis. The methodology specifically involves:

1) The provision for a given operational ensemble of a sample of filtered states whose stochastic errors are generally of reduced amplitude.
2. The selection of a subset of the filtered states to complement or supplant the operational ensemble members, under the constraint that the resulting distribution's mean and variance is comparable to that of the operational ensemble.

The hybrid ensemble that is the culmination of these two parts has mean and variance comparable to that of the operational ensemble by design. Thus, assuming that the hybrid also affords improved covariance, it should provide for multi-dimensional probability forecasts that are better than those based upon the operational ensemble.

3. Prototype filtering scheme

Investigation of the proposed methodology requires definition of a filtering scheme. The prototype filtering scheme adopted for the present analysis involves forming all possible pairs of operational members and then averaging the members in each pair. In the case of the National Center for Environmental Prediction (NCEP) Global Forecast System (GFS) 0000UTC initialization 11-member ensemble, there are 55 possible pairs and hence 55 so-called "pair-wise" filtered states.

Two observations together suggest that the pair-wise filtering might actually serve to reduce the amplitude of stochastic errors despite its simplicity. One is that the average of two dissimilar fields will generally have less amplitude than either one of the two fields. Evidence of this general rule can be seen in the simple schematics of Fig. 5. The second observation is that the stochastic-error fields in any two operational members are unlikely to be identical, given the origin of the errors in random processes. There are, however, no simple a priori observations to suggest that the pair-wise filtering can achieve an effective compromise between the elimination of errors and the smoothing of true flow components.

The choice of prototype filtering scheme prompts some additional remarks. For instance, it is easily shown that the mean of the entire set of pair-wise filtered states is identical to that of the operational ensemble. Similarly, it is easily shown that the mean of the pair-wise filtered states' true components is identical to that of the operational ensemble's true components, and that the mean of the filtered states' stochastic errors is zero. Thus, the filtered states display the desired mean properties outlined in Section 2. There is strong reason to believe that the membership of a given hybrid ensemble will exhibit close to the same mean properties. Consider that the members of the hybrid will, by design, be selected to optimize the hybrid's variance, and so are apt to be as dissimilar as possible. Since an appreciable change in mean properties would require inherent commonality, not dissimilarity, among the selected members, such a change is unlikely to be engendered by the construction of a hybrid ensemble.

Another remark is that averaging of ensemble output has proven most effective at producing reduced-error global patterns in synoptic-type fields such as 500 hPa geopotential height. With this point in mind, the pair-wise filtering scheme is applied only to fields of 500 hPa geopotential height in the present analysis. Also, the filtering scheme's application is focused on forecasts in the middle to late stages of the medium range, where forecasts of variables such as precipitation and surface temperature have little or no skill, and hence forecasts of synoptic flow patterns remain of considerable relevance.

Having defined a prototype filtering scheme, inquiry can now begin into several fundamental questions: Is the prototype filtering scheme able to produce on a systematic basis new flow states that are less encumbered by errors than the operational members are? Can hybrid ensembles actually offer improved covariance? Can hybrid ensembles yield multi-dimensional probabilistic forecasts that are better than those yielded by the operational ensemble? These questions are involved enough to be addressed separately, with the current paper's next section assigned to the first question and McLay and Martin (2005) dedicated to the latter two.

4. Evaluation of the prototype filtering scheme's efficacy

a. Data

Analysis is based upon 361 different National Center for Environmental Prediction (NCEP) Global Forecast System (GFS) 0000UTC initialization 11-member ensemble forecasts. These ensembles were generated during the one-year period between 21 December 2002 and 21 December 2003. The data were obtained on 2.5°-by-2.5° latitude-longitude grids in a cylindrical equidistant (CED) projection. As stated in Section 3, analysis is restricted to forecasts of 500 hPa geopotential height and the 192h forecast leadtime. The appropriate 0h leadtime control forecast was used as verification in all forecast error calculations. Several ensembles are missing or incomplete during the aforementioned one-year period and hence are unverifiable: the 12 July 2003, 27 August 2003, and 16 November 2003 ensembles. As a further consequence, 192h leadtime forecasts initialized 0000UTC 4 July 2003 and 19 August 2003 are unverifiable.

b. Designation of operational ensemble members and pair-wise filtered states

The 11 members of a given ensemble include the control, five members obtained through the addition to the control analysis of five different perturbations, and five members obtained through the subtraction from the control analysis of the five different perturbations. The operational designations for these members are $C002$, $P001$,...,$P005$, and $N001$,...,$N005$, respectively. For the present analysis, the 10 perturbed members ($P001$, $P002$, ..., $N001$, $N002$, ...) are referred to with alternative designations that reflect the members' rms distances from the ensemble mean. Specifically, the designations N_1, N_2,..., N_{10} are assigned to the member furthest from the mean, the member second furthest from the mean,....,the member closest to the mean, respectively. Note that N_1, N_2,..., N_{10} will not necessarily correspond to the same operationally designated member from ensemble to ensemble.

The designation for any specific pair-wise filtered state is simply the combination of its two component operational members' designations. For example, N_1N_2 refers to the pair-wise filtered state derived from operational members N_1 and N_2.

c. Definition of "Basic" and "Overall" Ensembles

Frequent reference is made in the analysis to the so-called "basic" and "overall" operational ensembles, and to the so-called "basic" and "overall" operational members. The "basic" operational ensemble is defined to consist of the control and the 10 perturbed members, and a "basic" member is a member of this ensemble. The "overall" operational ensemble is defined to consist of the basic operational ensemble plus the ensemble mean, and an "overall" member is a member of the overall ensemble.

d. Methodology

The reality is that for any given ensemble forecast only one true state is ultimately known to a good approximation, this being the forecast verification (i.e. the verifying analysis). An issue to be resolved is how it can be ascertained that the pair-wise filtered states are generally less corrupted by errors of stochastic origin than the operational members are, given just this one true state for reference. To facilitate this issue's resolution it is noted that the pair-wise filtered states have the same mean, and hence the same systematic error component, as the operational members. Thus, if in terms of some error measure the filtered states are found to be better approximations of true states than the operational members are, it is because the filtered states' stochastic error components have been reduced. Also, to facilitate resolution it is assumed that a suitable measure of the error in a given approximation of a true state is root-mean-square (rms) distance from that true state. Additionally, focus is placed on the lower bound of the range of rms error (rmse) of the overall operational ensemble and that of the set of pair-wise filtered states. The lower bound of the range of rmse for a given overall ensemble is determined by finding the minimum value of rmse associated with any member of the ensemble. The lower bound of the range of rmse for the associated set of filtered states is similarly determined. With these details for reference, it can be argued that the filtered states are generally better approximations of true states and hence less corrupted by stochastic errors if the lower bound of their range of rmse is smaller than the lower bound of the range of rmse of the overall ensemble on a systematic basis. To understand this, consider a scenario wherein one has a large sample of forecasts and, for 90% of the forecasts, the event (hereafter referred to as E_1) occurs that the filtered states' range of rmse has a smaller lower bound than that of the overall ensemble's range of rmse. Since E_1 occurs with high frequency across a large sample, it follows from the frequency interpretation of probability that E_1 is very likely to occur for any given forecast (Ross 1998). It also follows from elementary probability concepts that E_1 can be very likely to occur only if the following condition is satisfied: Each of a large majority of the possible true states $t_i, i = 1, ..., \infty$ associated with a given forecast is better approximated by some filtered state than by some overall operational ensemble member. Now, the question arises, 'Does this condition's satisfaction imply that most of the filtered states are better approximations of possible true states?'. Ostensibly, the answer to this is 'no'. One can imagine a scenario wherein the possible true states are all grouped near a relatively small proportion of the filtered states. In this scenario, the above condition could be satisfied, but only a minority of the filtered states might actually be better approximations of possible true states. However, such a scenario is contrived, and conceptualizations of it suggest that it would involve the operational ensemble being super-variant or extremely biased (Fig. 6). Neither of these circumstances frequently characterize operational ensembles, according to documented studies. Specifically, studies indicate that one of the major problems with operational ensembles is that they are too often sub-variant, not super-variant (e.g. Mylne et al. 2002). Also, observations regarding the 'proportion of outliers' (i.e. the proportion of time in which the verifying state lies outside the envelope of an ensemble distribution) for operational ensembles of 500 hPa geopotential height are not consistent with there frequently being extreme bias in these distributions. For instance, the results of Molteni et al. (1996) and Atger (1999) both indicate that the proportion of outliers for the European Centre for Medium-range Weather Forecasting (ECMWF) ensemble of 500 hPa height is less than 20% for typical medium-range forecast leadtimes. Were these distributions to be frequently characterized by extreme bias, one would expect the proportion of outliers to be considerably greater. In view of these points, the scenario wherein the possible true states are all grouped near a relatively small proportion of the filtered states should not be a common occurrence. This means that when the above condition (that ensures E_1 is very likely) is satisfied, it is generally because a large proportion (but not necessarily all) of the filtered states are better approximations of true states than are any members of the overall operational ensemble. It also means, by extension, that if E_1 is found to be very likely it is because most of the time a large proportion of the filtered states are better approximations of true states. With this conclusion, assessment of event E_1's likelihood (i.e. its frequency of occurrence) is the focal point of the relative comparison between the distribution of pair-wise filtered states and the operational ensemble.

e. Composite Relative Comparison

A relative comparison of the rmse characteristics of the pair-wise filtered state distribution and those of the operational ensemble distribution can be carried out using the large ensemble dataset described in Section 4a. The first part of the comparison involves each distribution's composite range of rmse. Computation of the composite range of rmse for the basic ensemble first involves obtaining the range of rmse for each of the 361 basic ensembles in the dataset. The range of rmse for any one of these ensembles is determined by finding the minimum and maximum values of rmse associated with any member of the given ensemble. Once these two quantities are determined for all 361 ensembles in the dataset, each quantity is then averaged over the dataset to obtain the composite range of rmse for the basic ensemble. Analogous steps determine the composite range of rmse for the set of pair-wise filtered states. To complement the composite ranges of rmse for the basic ensemble and the set of filtered states, the composite lower bound of the overall ensemble's range of rmse is also obtained. The lower bound of the range of rmse for any one of the overall ensembles is determined by finding the minimum value of rmse associated with any member of the given ensemble. Once this quantity is determined for all 361 ensembles in the dataset, it is averaged over the dataset to obtain the composite lower bound of the overall ensemble's range of rmse. For additional reference, the composite median value of rmse is obtained for both the basic ensemble and the set of pair-wise filtered states. The composite results are displayed in Fig. 7, and readily afford several suggestions. For instance, the relative position of the tops of the two bar diagrams indicates that the maximum value of rmse in any set of filtered states tends to be considerably less than that in the corresponding basic ensemble. Also, the position of the bottoms of the two bar diagrams relative to each other and to the line representing the composite lower bound of the overall ensemble's range of rmse indicates that the minimum rmse in any set of filtered states tends to be less than that in both the basic and overall ensembles. Finally, the relative position of the two line segments representing the composite median values of rmse suggests that the median rmse within any set of filtered states tends to be considerably less than that in the basic ensemble. Thus, on a composite basis, the set of pair-wise filtered states exhibits lower values of rmse than the operational ensemble does for at least three quantities: maximum, minimum, and median rmse. The fact that the composite minimum rmse for the set of filtered states is less than the composite minimum rmse within the overall ensemble is particularly important, because it is preliminary indication that the crucial event E_1 of Section 4d frequently occurs and, by extension, that the pair-wise filtering process can consistently identify a set of states that are less corrupted by stochastic errors than the operational members are.

f. Daily Relative Comparison

Additional insight can be gained by evaluating the pair-wise filtered state distribution's associated rmse values in relation to those of the operational ensemble on an ensemble-by-ensemble basis. For the assessment that follows, all 55 pair-wise filtered states were generated for each of the 361 operational ensembles in the dataset, and the rmse of each filtered state (as well as each operational member) was subsequently calculated. Additionally, for each of the 361 ensembles the overall operational member associated with the smallest value of rmse (hereafter referred to as the "best" overall operational member) was identified, and its value of rmse was used as a benchmark for comparison. To begin with the assessment, Fig. 8 displays for each of the 361 ensembles the number of filtered states with less rmse than the best overall operational member. It is inferred right away that for a great majority of the ensembles a filtered state exists that outperforms the best overall member. Indeed, this is true for 332 (or 92%) of the 361 ensembles. What is more, usually there is more than one filtered state that outperforms the best overall member. Specifically, on average there are ≈ 4.6 filtered states with less rmse than the best overall member. In other words, this means that for any given ensemble one can expect to find four or five filtered states with less rmse than the best overall member.

It remains to be seen just how much improvement in rmse might be associated with any one of these filtered states. To address this, Fig. 9 displays for each of the 361 ensembles the percentage improvement, in terms of rmse and relative to the best overall operational member, associated with the filtered state with the lowest rmse (hereafter referred to as the "best" filtered state). Readily apparent in the figure is that the percentage improvement frequently is at least 5%, and sometimes exceeds 10%. The average percent improvement is 5.2%, the minimum percent improvement is -9.6% (attained in association with the 10 October 2003 ensemble), and the maximum percent improvement is 20.0% (attained in association with the 09 February 2003 ensemble). Figure 10 provides a precise breakdown of the number of best filtered states whose percentage improvement in rmse falls within a given range of values. It was stated previously that fully 332 of 361 best filtered states (or $\approx$ 92%) register some finite positive percentage improvement relative to the best overall ensemble member. Now it can also be inferred from Fig. 10 that 180 of 361 best filtered states (or $\approx$ 50%) register $\geq$ 5% improvement, and that 36 of 361 best filtered states (or $\approx$ 10%) register $\geq$ 10% improvement. Stated differently, these results indicate that about one of every two ensembles will have an associated best filtered state that affords $\geq$ 5% improvement in rmse relative to the best overall solution of the given ensemble, and that about one of every 10 ensembles will have an associated best filtered state that affords $\geq$ 10% relative improvement.

To reiterate, the collective results presented in Figs. 8-10 show that a vast majority of the time the distribution of pair-wise filtered states will contain multiple states that have less rmse than the best overall operational member, and that frequently the distribution contains at least one state whose rmse is a notable improvement relative to that of the best overall operational member. These results are essential because they demonstrate that the event E_1 of Section 4d occurs on a very consistent basis. In other words, in accordance with the argument presented in Section 4d, these results establish that the distribution of pair-wise filtered states does, in fact, contain states that are generally less corrupted by stochastic errors than any operational members are.

5. Performance differentials within the distribution of filtered states

To this point the concern has been with a relative comparison of the pair-wise filtered states' performance and that of the operational members. Further understanding of the filtering process ultimately can be gained by examining the question of whether certain filtered states systematically prove more effective than others. Evidence that suggests an affirmative answer to this question is, in fact, forthcoming from at least two different perspectives.

a. Probabilistic Perspective

A probabilistic modelling perspective provides insight into two variations on the above question: 1) Do certain subsets of the filtered states account for a disproportionate number of the states that yield the most notable (i.e $\geq 10\%$) relative improvement?, and 2) Do certain subsets of the filtered states account for a disproportionate number of the "best" filtered states (i.e the filtered states with the minimum rmse for any given ensemble)?

Regarding the first of these questions, consider that 36 of the 361 best filtered states in the current work afforded $\geq 10\%$ improvement over the best overall operational member, and that 33 of these 36 filtered states were identified with the subset comprised of filtered states based upon one or more of the top five basic members most distant from the ensemble mean (i.e. N_1, N_2, N_3, N_4, N_5), S_{top5}. Examples of states in this subset include N_1C_2, N_2N_4, N_4N_5, N_4N_{10}, N_3N_7, etc. Examples of states not in this subset include N_6N_7, N_8C_2, N_9N_{10}, etc. The question is whether one could simply attribute the identification of those 33 filtered states with S_{top5} to chance, given that a total of 260 of the 361 best filtered states were identified with S_{top5}. The procedure for answering this question is as follows. First, the variable n is defined to be the number of the 36 filtered states associated with $\geq$ 10% improvement that are identified with S_{top5} through chance. Then, the probability mass function of n is calculated, and this function is used to perform an inference test on whether the event that 33 of the filtered states are identified with S_{top5} can be attributed to chance or not. The null hypothesis of the inference is that the event is a manifestation of chance, and the significance level of the inference is determined directly from the probability mass function as the probability of identifying $n =\geq 33$ filtered states with S_{top5}. If this probability is less than some arbitrary threshold, say 1%, then the null hypothesis is rejected and the event is assumed not to be a manifestation of chance. Two small steps facilitate an understanding of where the expression for the probability mass function comes from. The first is to refer to each of the 36 filtered states that afforded $\geq$ 10% improvement as a "success", for ease of discussion. The second is to restate the question posed above as, "If one has a sequence of 361 elements comprised of 36 successes and 325 non-successes in an arbitrary order, what is the probability of finding n successes in an arbitrary sample of 260 of the 361 elements?" This probability is readily computed as

$$P(n) = \frac{\begin{pmatrix} 36 \\ n \end{pmatrix} \begin{pmatrix} 361-36 \\ 260-n \end{pmatrix}}{\begin{pmatrix} 361 \\ 260 \end{pmatrix}}$$

where

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{a!}{(a-b)!b!} \, .$$

The lefthand term in the numerator is the number of ways to sample n of the 36 successes when order of selection doesn't matter, the righthand term in the numerator is the number of ways to sample 260-n of the 325 non-successes when order of selection doesn't matter, and the term in the denominator is the number of ways to sample 260 of the total of 361 elements when order of selection doesn't matter. Given this expression for $P(n)$, the corresponding probability mass function is obtained by computing $P(n)$ for $n = 0, 1, ..., 36$. Carrying out the probability mass function calculations and performing the inference test, it is found that the probability of identifying with S_{top5} through mere chance $n =\geq 33$ filtered states associated with $\geq 10\%$ improvement is $2.7x10^{-3}$. Furthermore, the number of filtered states expected to be identified with S_{top5} is only 25.9. Thus, there is strong evidence that S_{top5} is what might be considered a "preferred source" of the filtered states associated with the most notable (i.e. $\geq 10\%$) improvement over the best overall member.

A similar analysis can be performed for the case of subset S_{top4}, comprised of filtered states based upon one or more of the top four basic members most distant from the ensemble mean (i.e. N_1, N_2, N_3, N_4). Consider that 30 of the 36 best filtered states that afforded $\geq$ 10% improvement over the best overall operational member were identified with S_{top4}, and that 223 of the 361 best filtered states were identified with S_{top4}. Carrying out the analysis with these numbers, it is found that the probability of identifying with S_{top4} through mere chance $n =\geq 30$ filtered states associated with $\geq 10\%$ improvement is $3.1x10^{-3}$. Furthermore, the number of filtered states expected to be identified with S_{top4} is only 22.2. Thus, as with S_{top5}, there is strong evidence that S_{top4} is what might be considered a "preferred source" of the filtered states associated with the most notable (i.e. $\geq 10\%$) improvement over the best overall operational member.

One additional observation makes the findings related to S_{top4} and S_{top5} more meaningful. This is that the pair-wise filtered states that comprise S_{top4} and S_{top5} are based upon operational members relatively distant from the ensemble mean, and as such represent combinations of members that are the most different from one another in a root-mean-square sense. Given this observation, the above findings carry the implication that pair-wise filtering is most effective at realizing relative improvement in rmse when the basic members of a given pair-wise filtered state are not very similar in pattern.

Regarding the second question above, it is readily determined that at least one subset of the filtered states does account for a disproportionate number of the "best" filtered states. Consider some subset S of the 55 filtered states, of size m. If each of the 55 filtered states for a given ensemble is equally likely to be the best-performing filtered state, then the probability that the best filtered state will be identified with subset S is $p = m/55$. Assuming for simplicity that the identity of the best filtered state for a given ensemble is independent of the identities of the best filtered states in other (past) ensembles, then the number of times in a sample of n ensembles that the best filtered state is identified with subset S can be modelled as a binomially distributed random variable with expected value $E = np = nm/55$. Now, consider in particular the subset comprised of the 10 filtered states that are based upon the control (N_1C_2, N_2C_2, N_3C_2,....,$N_{10}C_2$), S_{c2}. With the sample of $n = 361$ ensembles available for the current work, the number of times that the best filtered state is identified with S_{c2} is expected to be $E = 361{\cdot}10/55 \approx 66$. Upon inspection, the best filtered state is found in S_{c2} 91 times, or about 40% more often than might be expected. Furthermore, the probability of the best filtered state being identified with S_{c2} through mere chance 91 or more times is only $5.5x10^{-4}$. Thus, the inference is that random processes are highly unlikely to account for there being so many best filtered states identified with S_{c2}. It follows, then, that S_{c2} could be considered a so-called "preferred origin" of the best filtered state. Upon reflection, this finding should not be considered a surprise, as it probably is a manifestation of the fact that in terms of rmse the control forecast is better than the perturbed members on a systematic basis. Nonetheless, it is more evidence that there are discernible performance differences within the distribution of pair-wise filtered states itself.

b. Anomaly projection perspective

An alternative and entirely different perspective from that of probabilistic modelling also suggests that certain pair-wise filtered states systematically prove more effective than others. To understand the basis of this perspective, consider first each basic ensemble member's anomaly relative to the ensemble mean. For any basic member of a given ensemble, this anomaly is simply defined to be the field obtained by taking the gridpoint-by-gridpoint difference between the basic member and the ensemble mean. The anomaly projection of any two different anomaly fields can also be defined as

$$\mathbf{r_{AB}} \quad = \quad \frac{\langle\ \mathbf{A}, \mathbf{B}\ \rangle}{\|A\|\ \|B\|}\ ,$$

where the anomaly fields are assumed to be in vector form, $\langle\,\rangle$ denotes an inner product, and $\|$ denotes vector magnitude. The anomaly projection may assume any value in the spectrum -1 to 1, and measures the two anomaly fields' similarity in terms of both pattern and sign. The absolute value of the projection measures the similarity of the two anomaly fields' patterns: The larger this value is, the more similar the two fields' patterns are. The sign of the projection indicates whether the two anomaly fields tend to have the same or opposite sign: If the projection is positive (negative), then the two fields tend to have the same (opposite) sign.

Each pair-wise filtered state is associated with an anomaly projection, for the reason that each filtered state is based upon two different basic members and these two members' anomaly fields have a certain anomaly projection. This is a relevant fact because there are suggestions that the performance of each filtered state is to some extent a function of its anomaly projection value. To see where these suggestions arise requires the following analysis. First, for any given ensemble, the 55 filtered states are sorted from smallest to largest in terms of rmse. Next, the filtered state with the smallest rmse is assigned rank 1, the filtered state with the second smallest rmse is assigned rank 2, and so forth, until the filtered state with the largest error is assigned rank 55. Then, the anomaly projection corresponding to a given filtered state is assigned the same rank as the filtered state. That is, the anomaly projection corresponding to the filtered state with rank 1 is put in a rank 1 "bin", the anomaly projection corresponding to the filtered state with rank 2 is put in a rank 2 "bin", etc. The preceding steps are carried out for every available ensemble. At the culmination of this effort, a record is obtained of the anomaly projection values associated with each of the 55 ranks. In other words, with the sample of 361 ensembles available for this work, the 361 anomaly projection values associated with rank 1 are obtained, the 361 anomaly projection values associated with rank 2 are obtained, and so forth. This record can be used to assess whether specific forecast ranks of interest are characteristically populated by filtered states with a certain type of projection value.

For the current work, the record was tabulated using the entire sample of 361 ensembles. Figure 11 displays some standard statistics of the projection values associated with each of the 55 ranks, as derived from this record. First, consider the ranks associated with the worst rmse, ranks 45-55. Both the mean and median anomaly projection values for each of these ranks are substantially more positive than the overall mean projection value, and both the mean and median projection values for each of the ranks 50-55 are positive. The standard deviation of projection value for each of the forecast ranks 45-55 is also relatively low, indicating that the anomaly projections for these ranks do not tend to depart far into the negative part of the spectrum. To add emphasis to these points, it is found that fully 329 (or $\approx$ 91%) of the 361 anomaly projection values associated with rank 55 are positive (not shown). Thus, there are clear indications that the ranks associated with the worst rmse (i.e. ranks 45-55) tend to be populated by filtered states with anomaly projections in or very near the positive part of the spectrum. Consider also the ranks associated with the best rmse, ranks 1-5. Both the mean and median anomaly projection values for each of these ranks are negative, indicating that there is some inclination for these ranks to be populated by filtered states with negative anomaly projections. Indeed, it is found that 245 (or $\approx$ two-thirds) of the 361 anomaly projection values associated with rank 1 are negative. However, Fig. 11 also reveals that the mean anomaly projection value for each of the ranks 1-5 is quite close to the overall mean projection value, rather than substantially less than it. In addition, the standard deviation of projection value for each of the ranks is relatively high, indicating that the anomaly projections for each rank generally have anywhere from large negative to small positive values. Thus, it can be inferred from the above observations that the ranks associated with the best rmse are, for the greater proportion of time, populated by filtered states with negative anomaly projection values, and are populated for a lesser but not insignificant proportion of time by filtered states with relatively small positive anomaly projection values. It can also be inferred that filtered states with moderate to large positive anomaly projections do not populate these ranks with any substantial frequency. Collectively, the observations from Fig. 11 regarding ranks 1-5 and 45-55 suggest that filtered states with moderate or greater positive anomaly projection values tend not to perform as well as filtered states with negative or small positive anomaly projection values. There is also some suggestion that filtered states with positive projection values tend to perform worse than filtered states with negative projection values. This last suggestion is bolstered by observations of the anomaly projections associated with those best-performing filtered states that afford $\geq 10\%$ relative improvement. In 31 of the 36 cases in which the best filtered state affords $\geq 10\%$ improvement over the best overall operational member, the best filtered state is associated with a negative anomaly projection. This is $\approx 86\%$ of the time, clearly the much greater proportion. In 115 of the 124 cases in which the best filtered state affords $\geq 10\%$ improvement over the best basic ensemble member, the best filtered state is associated with a negative anomaly projection. This is $\approx 93\%$ of the time, again clearly the much greater proportion. Additional analysis (not shown) reveals one other suggestion. In particular, examination of the projections that populate rank 1 indicates that these projections are not generally any of those with the largest values (either positive or negative) in any given set of projection values. Thus, while pair-wise filtered states that have large positive projection values tend to perform the worst, pair-wise filtered states that have large negative projection values do not tend to perform the best. Taking this suggestion together with the other findings, the overall impression is that the best performing pair-wise filtered states tend to be those with moderate to small negative anomaly projection values. The practical interpretation that follows is that the pair-wise filtering process is most effective when the two basic members of a filtered state have anomaly fields that are only modestly similar in pattern and that are of generally opposite sign.

6. Discussion and Conclusions

This paper has two main objectives. One is to introduce an ensemble post-processing methodology designed to counter the effect of stochastic model errors on ensemble covariance. The first step in the methodology is to perform a series of filtering experiments on the operational ensemble members, with the aim of obtaining a set of states with reduced-amplitude stochastic error components. Since inquiry into the methodology requires that some form of filtering scheme be established, the other main objective of this paper is to define and evaluate a prototype filtering scheme. This scheme, referred to as "pair-wise" filtering, involves forming all possible pairs of operational members and then averaging the members in each pair. In the evaluation of the scheme, emphasis is placed on the outstanding issue of whether the pair-wise filtered states are, in general, less corrupted by errors of stochastic origin than the operational members are. It is argued, based upon elementary probabilistic concepts, that the pair-wise filtered states must generally be less corrupted if the lower bound of their range of rmse is very consistently smaller than the lower bound of the overall ensemble's range of rmse. With this idea as a guideline, 361 different NCEP GFS 0000UTC initialization 2.5°-by-2.5° resolution 11-member ensemble forecasts of 192h leadtime 500 hPa geopotential height are analyzed. This large sample of ensembles spans the period 0000UTC 21 December 2002 to 0000UTC 21 December 2003. All (55) possible pair-wise filtered states are derived for each of the 361 ensembles, and the performance of each set of filtered states is subsequently verified. Results of the analysis show that for fully 332 (or 92%) of the 361 ensembles examined the set of pair-wise filtered states contains at least one state that outperforms the best overall operational member (including the control, the perturbed members, and the ensemble mean) in terms of rmse. In fact, for each of the 361 ensembles there are, on average, 4-5 filtered states that have lower rmse than the best overall operational member. Further analysis shows that it is not uncommon for the set of filtered states to include at least one state which affords what might be termed noteworthy improvement (i.e. improvement of 5% or more) in rmse relative to the best overall operational member. Specifically, about one of every two sets of filtered states yields such a state. Additionally, about one of every 10 sets of filtered states yields a state that affords $\geq$ 10% improvement in rmse relative to the best overall operational member. The maximum observed improvement of a filtered state relative to the best overall operational member is 20.0%. To summarize, the analysis of the 361 ensembles indicates that a vast majority of the time the set of pair-wise filtered states contains multiple states that have less rmse than the best overall operational member, and also that the set often contains at least one state whose rmse is a noteworthy improvement relative to that of the best overall operational member.

Recalling the arguments of Section 4d, the above findings support the conclusion that for any given ensemble forecast the pair-wise filtered states are, in general, less corrupted by errors of stochastic origin than the operational members are. These findings affirm the utility of the simple pair-wise filtering procedure and hence encourage preliminary research regards the second part of the proposed post-processing methodology, which is the generation of a hybrid ensemble using some subset of the filtered states. Whether or not a hybrid ensemble can exhibit improved covariance and provide better multi-dimensional probabilistic forecasts is the issue that now takes precedence. McLay and Martin (2005) investigate the hybrid ensemble generation problem via idealized experiments and probabilistic forecast evaluations.

Some attention is also devoted in the present paper to the question of whether certain pair-wise filtered states are systematically more effective than others. Indeed, through probabilistic modelling, two subsets emerge as being preferred sources of those filtered states that yield the most noteworthy relative improvement in rmse ($\geq$ 10%). These subsets include the one comprised of filtered states based upon one or more of the top four operational members most distant from the ensemble mean (S_{top4}), and the one comprised of filtered states based upon one or more of the top five operational members most distant from the ensemble mean (S_{top5}). These subsets are "preferred" sources in the sense that the filtered states associated with the most notable relative improvement originate from these subsets significantly more often than can be expected through mere chance. The findings related to S_{top4} and S_{top5} are interesting when it is considered that the pair-wise filtered states that comprise S_{top4} and S_{top5} are based upon operational members relatively distant from the ensemble mean, and as such represent combinations of members that are the most different from one another in a root-mean-square sense. The implication, then, is that pair-wise filtering is most effective at eliminating stochastic error components when the two operational members of a given filtered state are fairly dissimilar. It could be reasoned that this is to be expected, on the grounds that the more dissimilar two members are, the more likely it is that the stochastic errors in the two members are different and, therefore, the more likely it is that the averaging process will produce a state with diminished stochastic errors.

Other suggestions of systematic performance differentials are gained by assessing the pair-wise filtered states' rmse performance as a function of so-called anomaly projection value. Each filtered state in a given set is associated with an anomaly projection value, because each of the two basic members upon which the filtered state is based is associated with an anomaly field relative to the ensemble mean and the two basic members' anomaly fields have a certain projection onto one another. Two suggestions are engendered by the anomaly projection value analysis. One is that filtered states with moderate or greater positive anomaly projection values tend not to perform as well as filtered states with negative or small positive anomaly projection values. The other is that filtered states with negative anomaly projection values tend to perform better than filtered states with positive anomaly projection values, though not overwhelmingly so. The practical interpretation of these two suggestions is that pair-wise filtering will be most effective when the operational members of a given pair have anomaly fields that are both somewhat similar in pattern and generally of opposite sign. In such a case there will be some cancellation of anomaly field components during the averaging process, but the cancellation will fall short of being wholesale. This will allow for some stochastic error components to be eliminated, but at the same time will ensure that a fair amount of any legitimate information found in the anomaly fields will be retained. The anomaly projection results are consistent with those gained from the probabilistic modelling, in the sense that they also indicate that pair-wise filtered states are most effective when based upon operational members that don't have great similarities in pattern.

It is noted, finally, that while the systematic performance differentials highlighted via the probabilistic modelling and anomaly projection analysis help to reveal the circumstances in which pair-wise filtering is most effective, they may also be of practical consequence. This is for the reason that knowledge of the performance differentials might be exploited in efforts to select subsets of the pair-wise filtered states for incorporation into hybrid ensembles.

Acknowledgments

This research was supported by a grant from the University of Wisconsin-Madison.

References

Alhamed, A., S. Lakshmivarahan, and D. J. Stensrud, 2002: Cluster analysis of multimodel ensemble data from SAMEX. *Mon. Wea. Rev.*, **130**, 226-256.

Atger, F., 1999: Tubing: An alternative to clustering for the classification of ensemble forecasts. *Wea. Forecasting*, **14**, 741-757.

Barkmeijer, J., T. Iversen, and T. N. Palmer, 2003: Forcing singular vectors and other sensitive model structures. *Quart. J. Roy. Meteor. Soc.*, **129**, 2401-2423.

Bright, D. R., and S. L. Mullen, 2002: Short-range ensemble forecasts of precipitation during the southwest monsoon. *Wea. Forecasting*, **17**, 1080-1100.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887-2908.

Colucci, S. J., and D. P. Baumhefner, 1998: Numerical prediction of the onset of blocking: A case study with forecast ensembles. *Mon. Wea. Rev.*, **126**, 773-784.

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132-1147.

Evans, R. E., M. S. J. Harrison, R. J. Graham, and K. R. Mylne, 2000: Joint medium-range ensembles from the Met. Office and ECMWF systems. *Mon. Wea. Rev.*, **128**, 3104-3127.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550-560.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711-724.

Harrison, M. S. J., T. N. Palmer, D. S. Richardson, and R. Buizza, 1999: Analysis and model dependencies in medium-range ensembles: Two transplant case studies. *Quart. J. Roy. Meteor. Soc.*, **125**, 2487-2515.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225-1242.

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

McLay, J. G., and J. E. Martin, 2005: Using filtering to mitigate stochastic model errors' effect on ensemble covariance. Part II: Use of filtered states in hybrid ensembles. *Mon. Wea. Rev.*, submitted.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Mylne, K. R., R. E. Evans, and R. T. Clark, 2002: Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Quart. J. Roy. Meteor. Soc.*, **128**, 361-384.

Orrell, D., 2003: Model error and predictability over different timescales in the Lorenz '96 systems. *J. Atmos. Sci.*, **60**, 2219-2228.

Orrell, D., L. Smith, J. Barkmeijer, and T. N. Palmer, 2001: Model error in weather forecasting. *Nonlinear Proc. Geoph.*, **8**, 357-371.

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, **63**, 71-116.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models. *Quart. J. Roy. Meteor. Soc.*, **127**, 279-304.

Palmer, T. N., R. Buizza, and F. Lalaurette, 1999: Performance of the ECMWF ensemble prediction system. *Proc. ECMWF Workshop on Predictability*, ECMWF, Reading, United Kingdom, 203-214. [Available from ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, United Kingdom.]

Ross, S. M., 1998: *A First Course in Probability*. Prentice-Hall, 514 pp.

Roulston, M. S., and L. A. Smith, 2003: Combining statistical and dynamical ensembles. *Tellus*, **55A**, 16-30.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433-446.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial-condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077-2107.

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

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729-747.

Wang, X., and C. H. Bishop, 2004: Ensemble augmentation with a new dressing kernel. Preprints, *20th conference on weather analysis and forecasting/16th conference on numerical weather prediction,* Seattle, WA, Amer. Meteor. Soc., J6.4.

Figure Captions

Figure 1. Conceptual depiction of forecast trajectories evolved with the dynamics of the real atmosphere and of a numerical model. The light grey striped shading represents an initial PDF, and the grey dot represents an initial state sampled from the PDF. Black solid arrow (dashed arrow) is the forecast trajectory evolved from the initial state using the dynamics of the real atmosphere (of a numerical model). Time increases from left to right, as indicated by the axis at the bottom of the figure.

Figure 2. Conceptual depiction of errors introduced by a numerical model. Same as for Fig. 1, except that the open circle (black dot) represents the state at time t on the forecast trajectory evolved with the dynamics of the real atmosphere (of a numerical model). The difference between the two states, represented as the thick solid grey arrow, defines the errors introduced into the forecast as of time t by the numerical model.

Figure 3. Illustration of the nature of the error component associated with systematic processes. The box encompasses a two-dimensional forecast sample space for some time t. Each open circle m_i (black dot n_i) represents the forecast state obtained by evolving some initial state s_i forward in time using the dynamics of the real atmosphere (of a numerical model). A grey arrow represents the error associated with each n_i.

Figure 4. Illustration of the nature of the error component associated with stochastic processes. Symbology the same as for Figure 3.

Figure 5. Two simple illustrations of the effect of averaging two dissimilar vectors. Each panel represents a two-dimensional space, with coordinate axes given as thin solid lines. In each panel, two dissimilar vectors are represented as solid black arrows. The average of the two dissimilar vectors is represented as a dotted black arrow. Note that the average vector has smaller magnitude than either of the two vectors it derives from.

Figure 6. Several conceptualizations of the scenario wherein the possible true states are all grouped near a small proportion of the filtered states. Each panel encompasses the sample space for some ensemble forecast. In each panel the forecast state associated with a given ensemble member is represented as a large dot, and the pair-wise samples associated with each pair of ensemble members are represented as X's. The continuum of possible true states is defined with grey shading. In each of these examples the event E_1 is very likely to occur just because the possible true states are all grouped near the pair-wise samples with asterisks.

Figure 7. Composite range of 192h 500 hPa geopotential height rmse (m) for a) the basic operational ensemble and b) the distribution of pair-wise filtered states. In each case the composite range of rmse is delimited by a grey shaded box and the composite median rmse is depicted with a dashed line. The composite lower bound of the overall operational ensemble's range of rmse is depicted with the dotted line.

Figure 8. Number of pair-wise filtered states whose 192h 500 hPa geopotential height rmse is less than that of the best overall operational member, for each of the 361

ensembles available between 21 December 2002 and 21 December 2003. Each bar represents the number of filtered states for a specific ensemble. The bars are arranged in chronological order, beginning with the bar for the 21 December 2002 ensemble in the upper left corner and proceeding left-to-right and top-to-bottom. The labels ‘J’, ‘F’, ‘M’, etc. identify the bars associated with the ensembles on the first day of January, February, March, and so forth. The number of filtered states associated with a given bar is determined by the scale on the ordinate of each rectangular plot. The bold horizontal line indicates the average over all 361 ensembles of the number of pair-wise filtered states whose rmse is less than that of the best overall operational member.

Figure 9. The best pair-wise filtered state’s percentage improvement in 192h 500 hPa geopotential height rmse relative to the rmse of the best overall operational member, for each of the 361 ensembles available between 21 December 2002 and 21 December 2003. Each bar represents the percentage improvement for a specific ensemble. The percentage improvement associated with a given bar is indicated by the scale on the ordinate of each rectangular plot. The bold horizontal line indicates the average over all 361 ensembles of the best filtered state’s percentage improvement in rmse. Layout otherwise the same as for Figure 8.

Figure 10. Number of best filtered states whose percentage improvement, α, in 192h 500 hPa geopotential height rmse relative to the rmse of the best overall operational member lies within a specified range. The abscissa indicates the number of best filtered states. The ranges of percentage relative improvement are given along the ordinate. Bars with

light (dark) grey shading are associated with best filtered states that yield positive (negative) percentage relative improvement.

Figure 11. Mean and median anomaly projection value, and mean anomaly projection value plus or minus one standard deviation, as a function of forecast rank. Forecast rank is given on the abscissa and anomaly projection value is given on the ordinate. The “+” (“•”) symbols indicate the mean (median) anomaly projection value associated with each forecast rank. The upper (lower) sequence of diamond symbols denotes for each forecast rank the mean anomaly projection value plus (minus) one standard deviation. The grey horizontal dashed line indicates the mean anomaly projection value for all forecast ranks.

Figure 1. Conceptual depiction of forecast trajectories evolved with the dynamics of the real atmosphere and of a numerical model. The light grey striped shading represents an initial PDF, and the grey dot represents an initial state sampled from the PDF. Black solid arrow (dashed arrow) is the forecast trajectory evolved from the initial state using the dynamics of the real atmosphere (of a numerical model). Time increases from left to right, as indicated by the axis at the bottom of the figure.

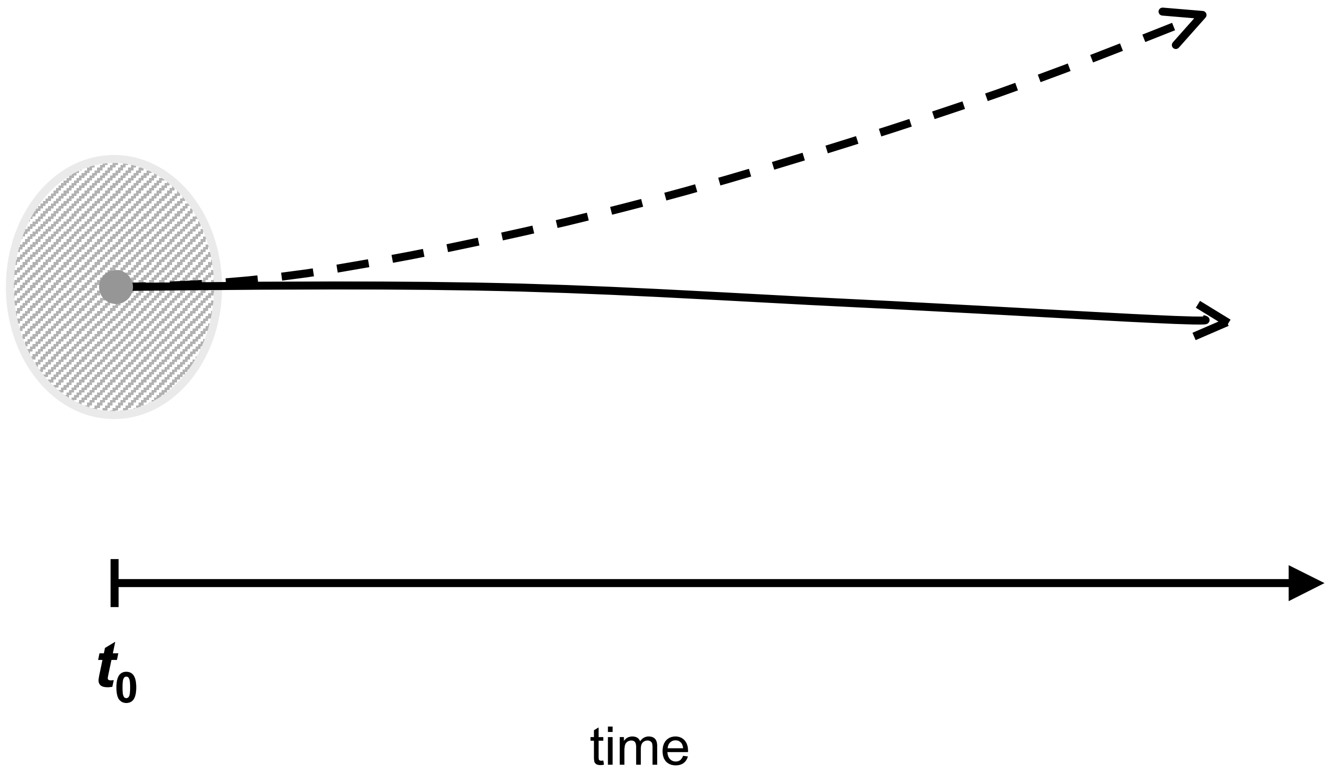


Figure 2. Conceptual depiction of errors introduced by a numerical model. Same as for Fig. 1, except that the open circle (black dot) represents the state at time *t* on the forecast trajectory evolved with the dynamics of the real atmosphere (of a numerical model). The difference between the two states, represented as the thick solid grey arrow, defines the errors introduced into the forecast as of time t by the numerical model.

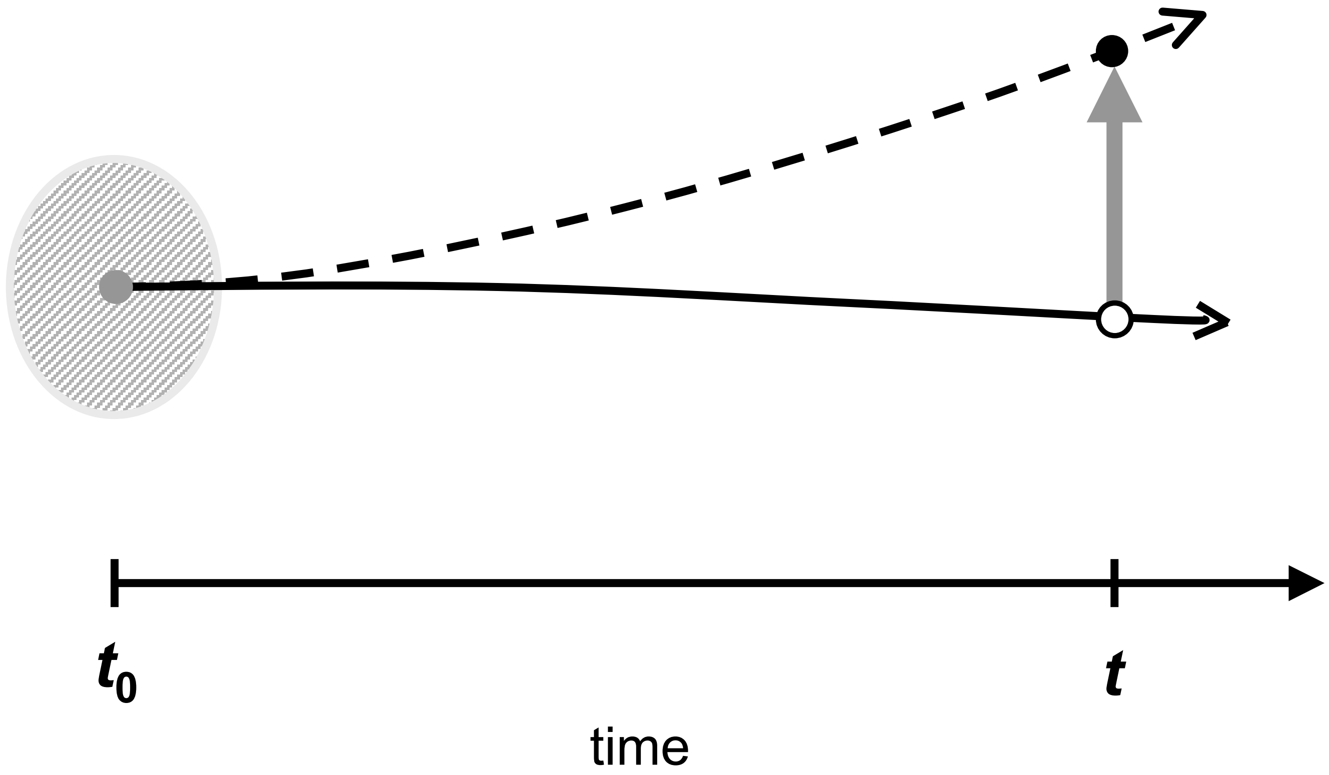


Figure 3. Illustration of the nature of the error component associated with systematic processes. The box encompasses a two-dimensional forecast sample space for some time t. Each open circle m_i (black dot n_i) represents the forecast state obtained by evolving some initial state s_i forward in time using the dynamics of the real atmosphere (of a numerical model). A grey arrow represents the error associated with each n_i.

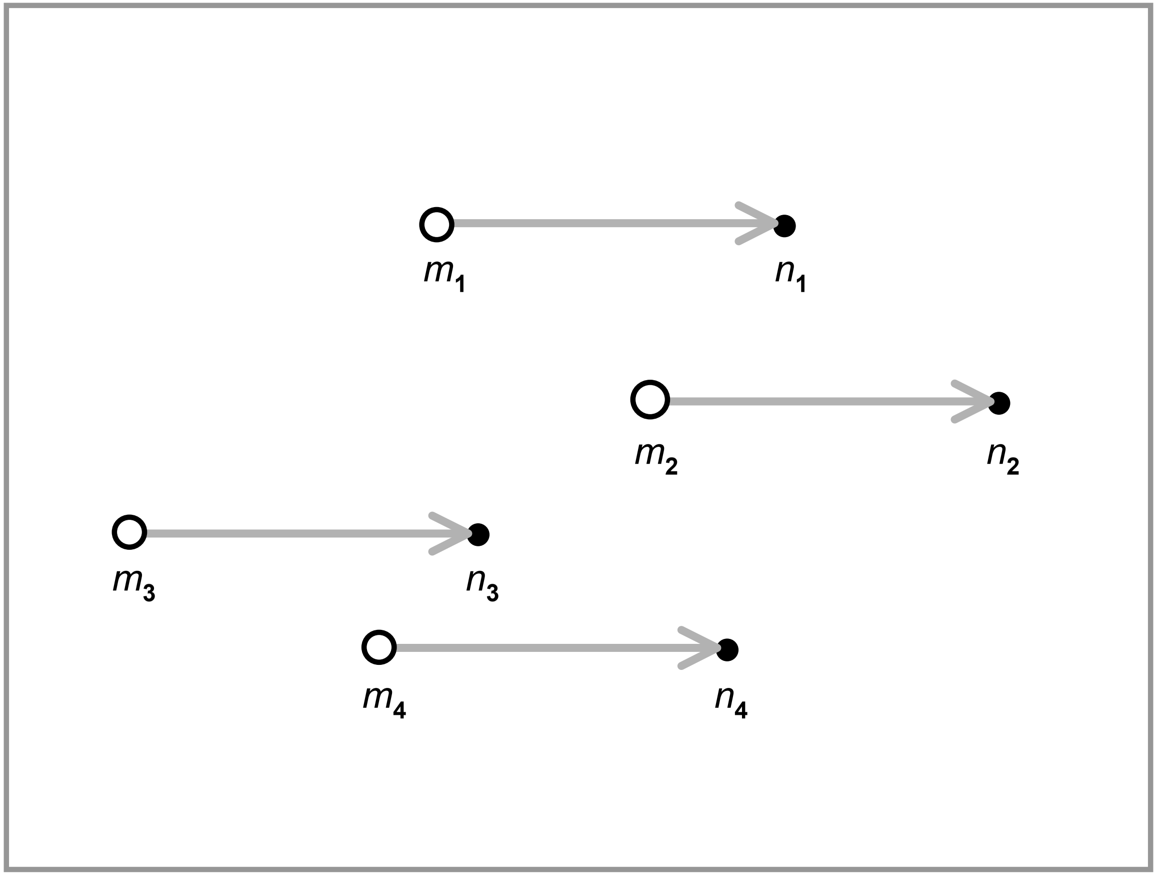


Figure 4. Illustration of the nature of the error component associated with stochastic processes. Symbology the same as for Figure 3.

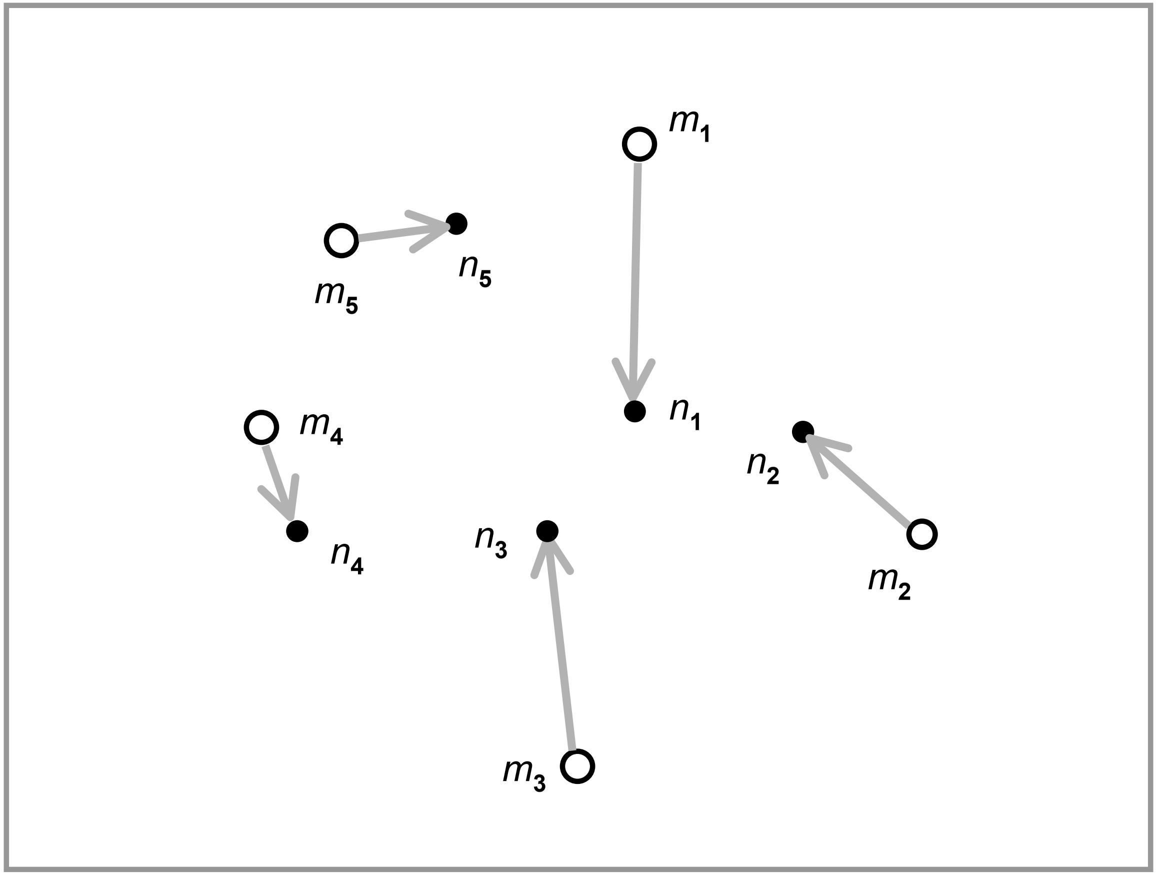


Figure 5. Two simple illustrations of the effect of averaging two dissimilar vectors. Each panel represents a two-dimensional space, with coordinate axes given as thin solid lines. In each panel, two dissimilar vectors are represented as solid black arrows. The average of the two dissimilar vectors is represented as a dotted black arrow. Note that the average vector has smaller magnitude than either of the two vectors it derives from.

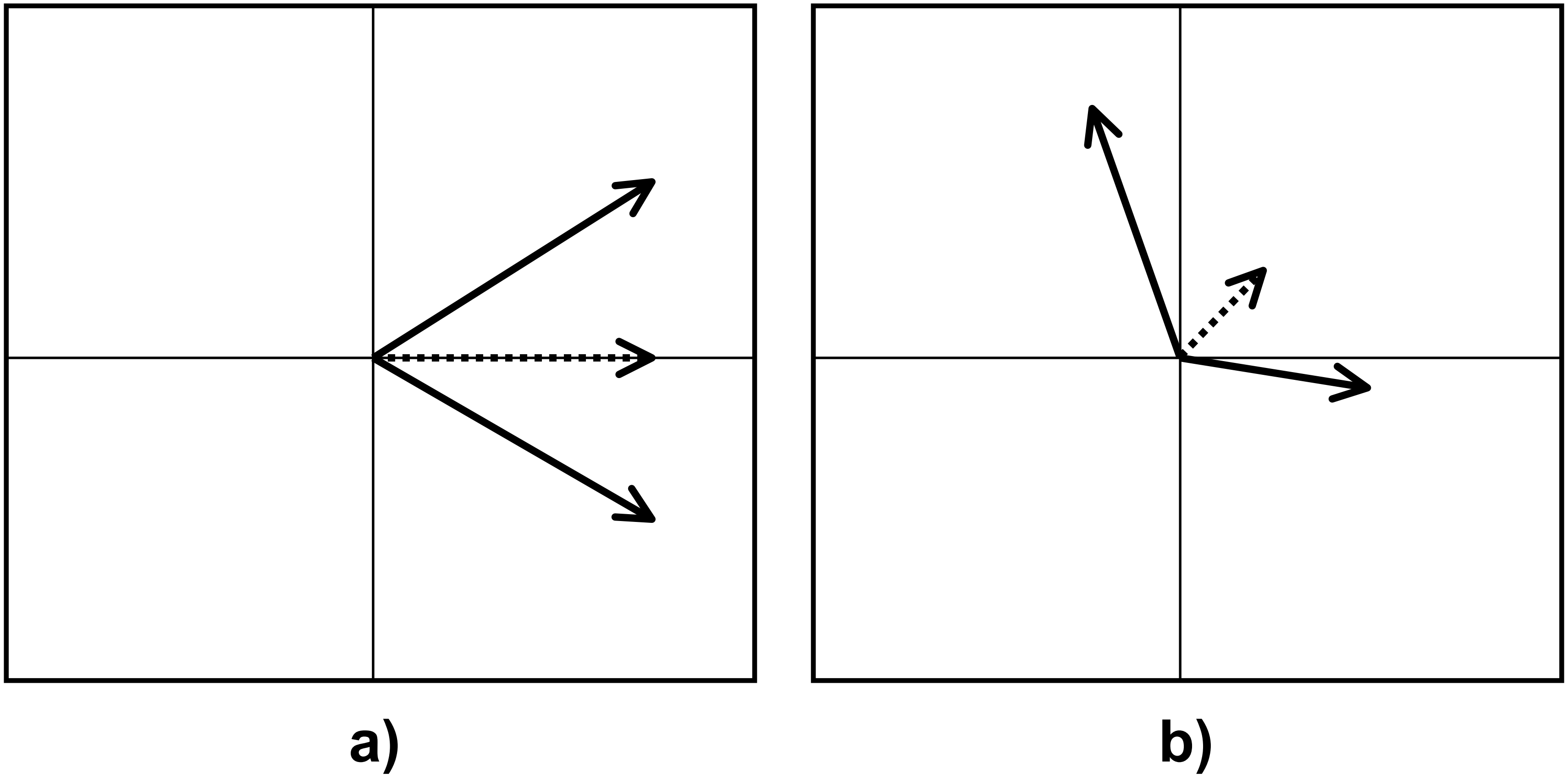


Figure 6. Several conceptualizations of the scenario wherein the possible true states are all grouped near a small proportion of the filtered states. Each panel encompasses the sample space for some ensemble forecast. In each panel the forecast state associated with a given ensemble member is represented as a large dot, and the pair-wise samples associated with each pair of ensemble members are represented as X's. The continuum of possible true states is defined with grey shading. In each of these examples the event E_1 is very likely to occur just because the possible true states are all grouped near the pair-wise samples with asterisks.

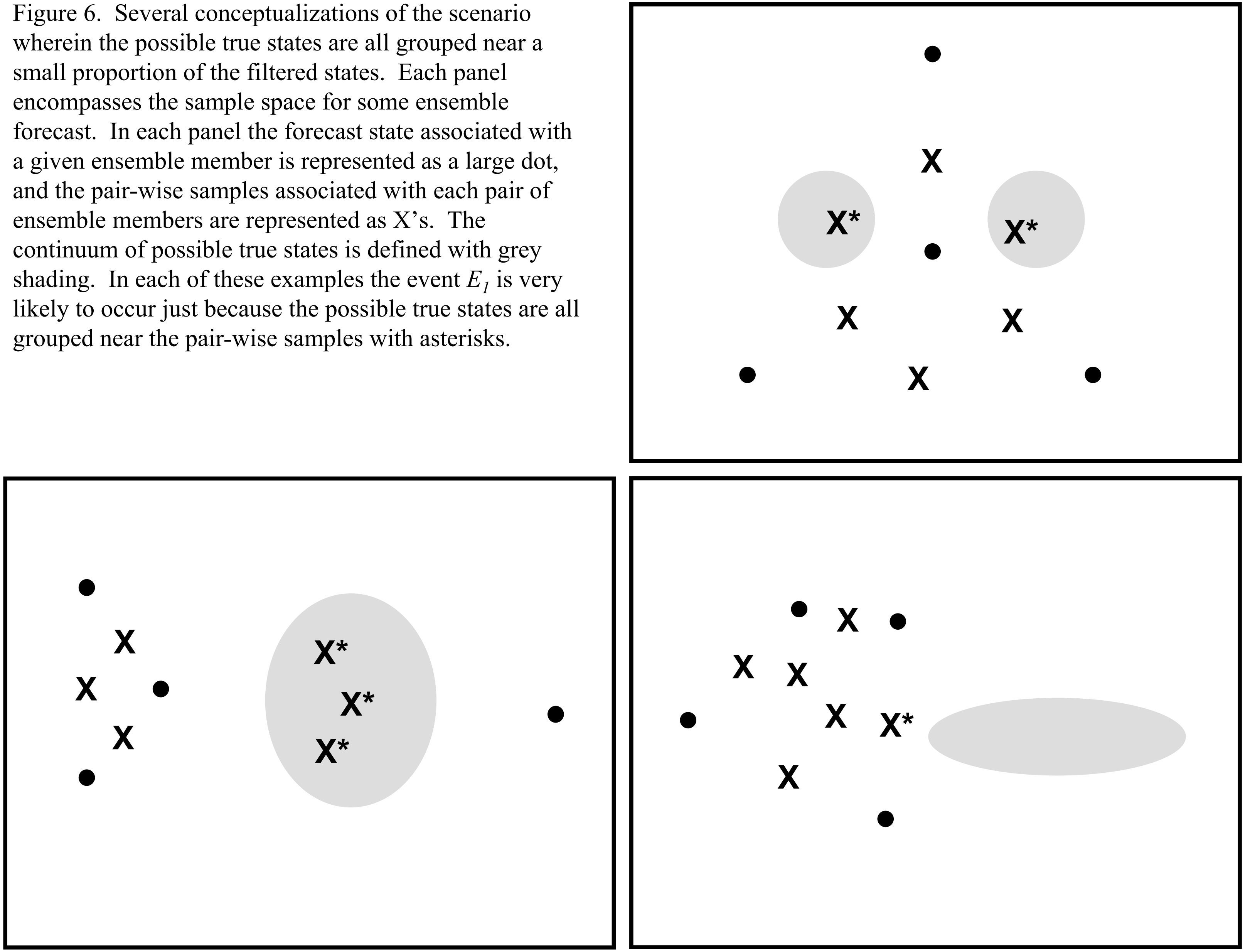


Figure 7. Composite range of 192h 500 hPa geopotential height rmse (m) for a) the basic operational ensemble and b) the distribution of pair-wise filtered states. In each case the composite range of rmse is delimited by a grey shaded box and the composite median rmse is depicted with a dashed line. The composite lower bound of the overall operational ensemble's range of rmse is depicted with the dotted line.

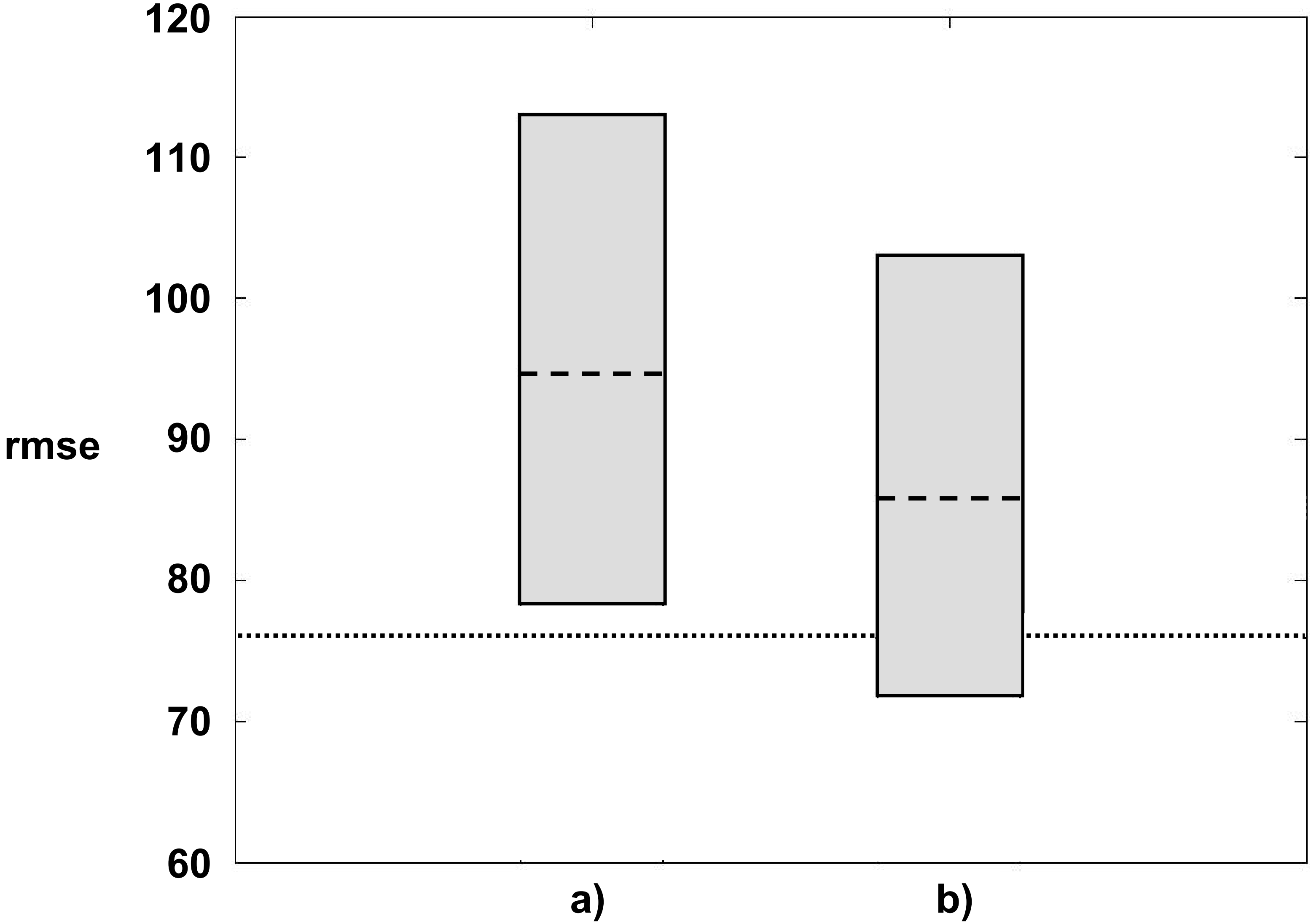


Figure 8. Number of pair-wise filtered states whose 192h 500 hPa geopotential height rmse is less than that of the best overall operational member, for each of the 361 ensembles available between 21 December 2002 and 21 December 2003. Each bar represents the number of filtered states for a specific ensemble. The bars are arranged in chronological order, beginning with the bar for the 21 December 2002 ensemble in the upper left corner and proceeding left-to-right and top-to-bottom. The labels ‘J’, ‘F’, ‘M’, etc. identify the bars associated with the ensembles on the first day of January, February, March, and so forth. The number of filtered states associated with a given bar is determined by the scale on the ordinate of each rectangular plot. The bold horizontal line indicates the average over all 361 ensembles of the number of pair-wise filtered states whose rmse is less than that of the best overall operational member.

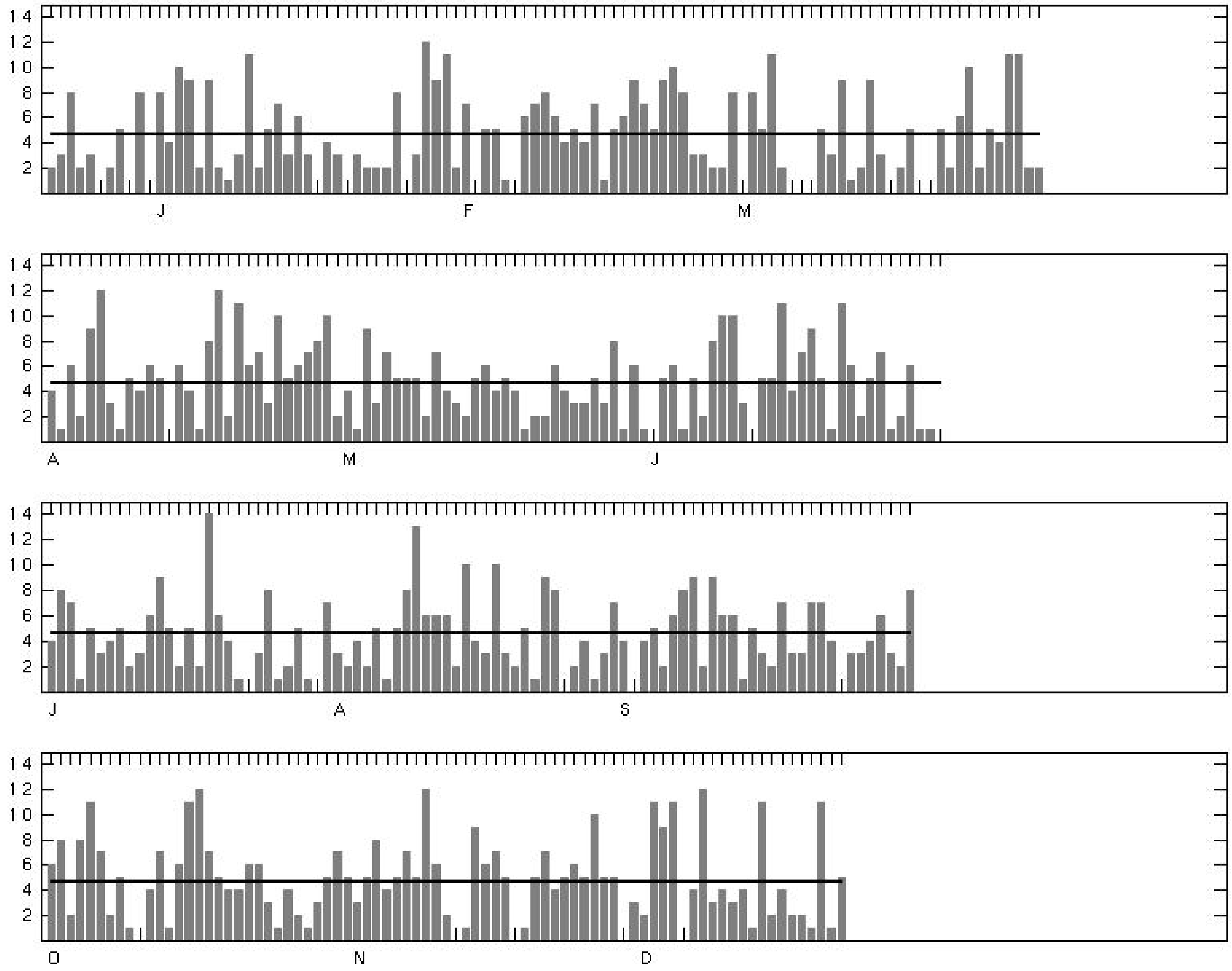


Figure 9. The best pair-wise filtered state's percentage improvement in 192h 500 hPa geopotential height rmse relative to the rmse of the best overall operational member, for each of the 361 ensembles available between 21 December 2002 and 21 December 2003. Each bar represents the percentage improvement for a specific ensemble. The percentage improvement associated with a given bar is indicated by the scale on the ordinate of each rectangular plot. The bold horizontal line indicates the average over all 361 ensembles of the best filtered state's percentage improvement in rmse. Layout otherwise the same as for Figure 8.

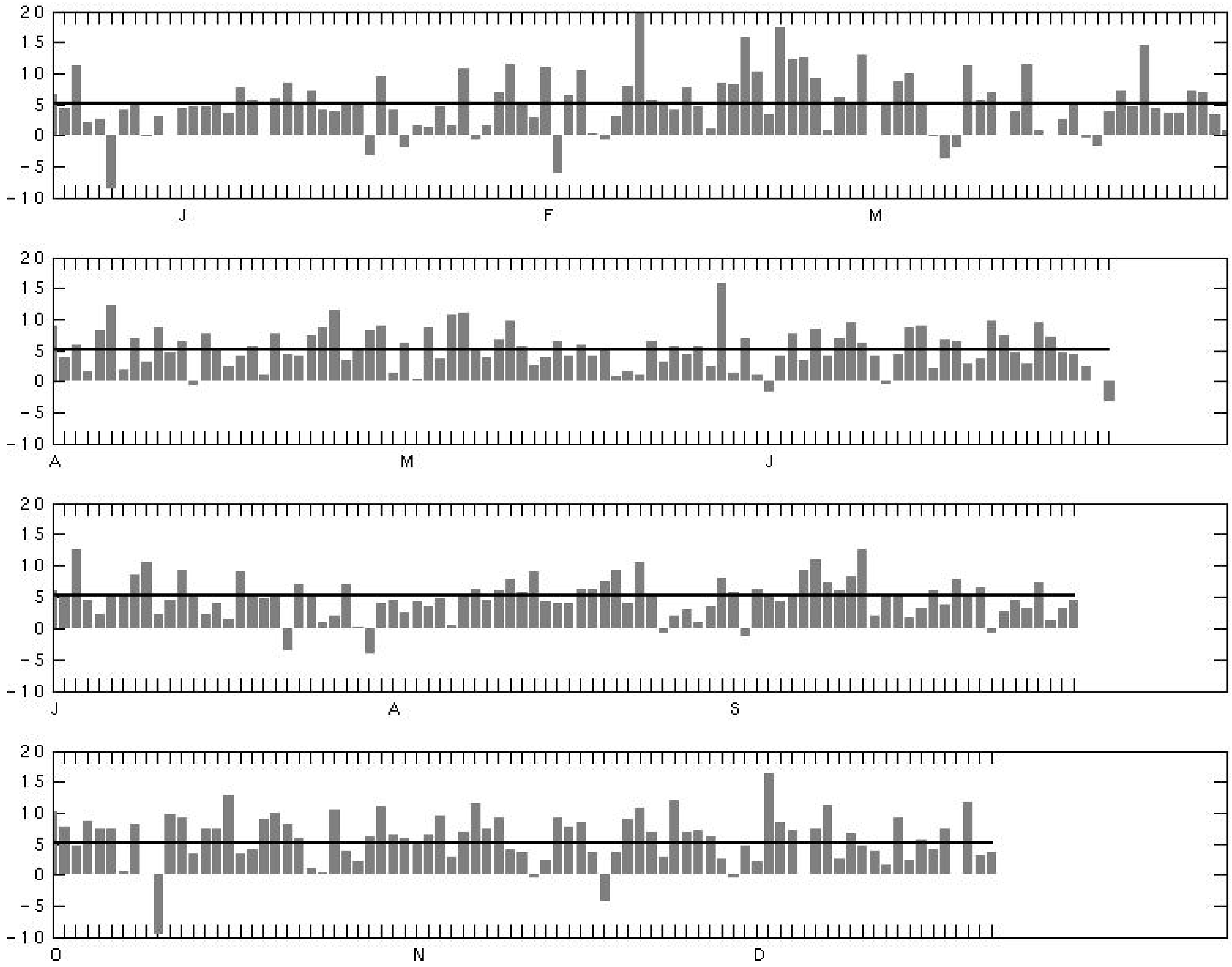


Figure 10. Number of best filtered states whose percentage improvement, α, in 192h 500 hPa geopotential height rmse relative to the rmse of the best overall operational member lies within a specified range. The abscissa indicates the number of best filtered states. The ranges of percentage relative improvement are given along the ordinate. Bars with light (dark) grey shading are associated with best filtered states that yield positive (negative) percentage relative improvement.

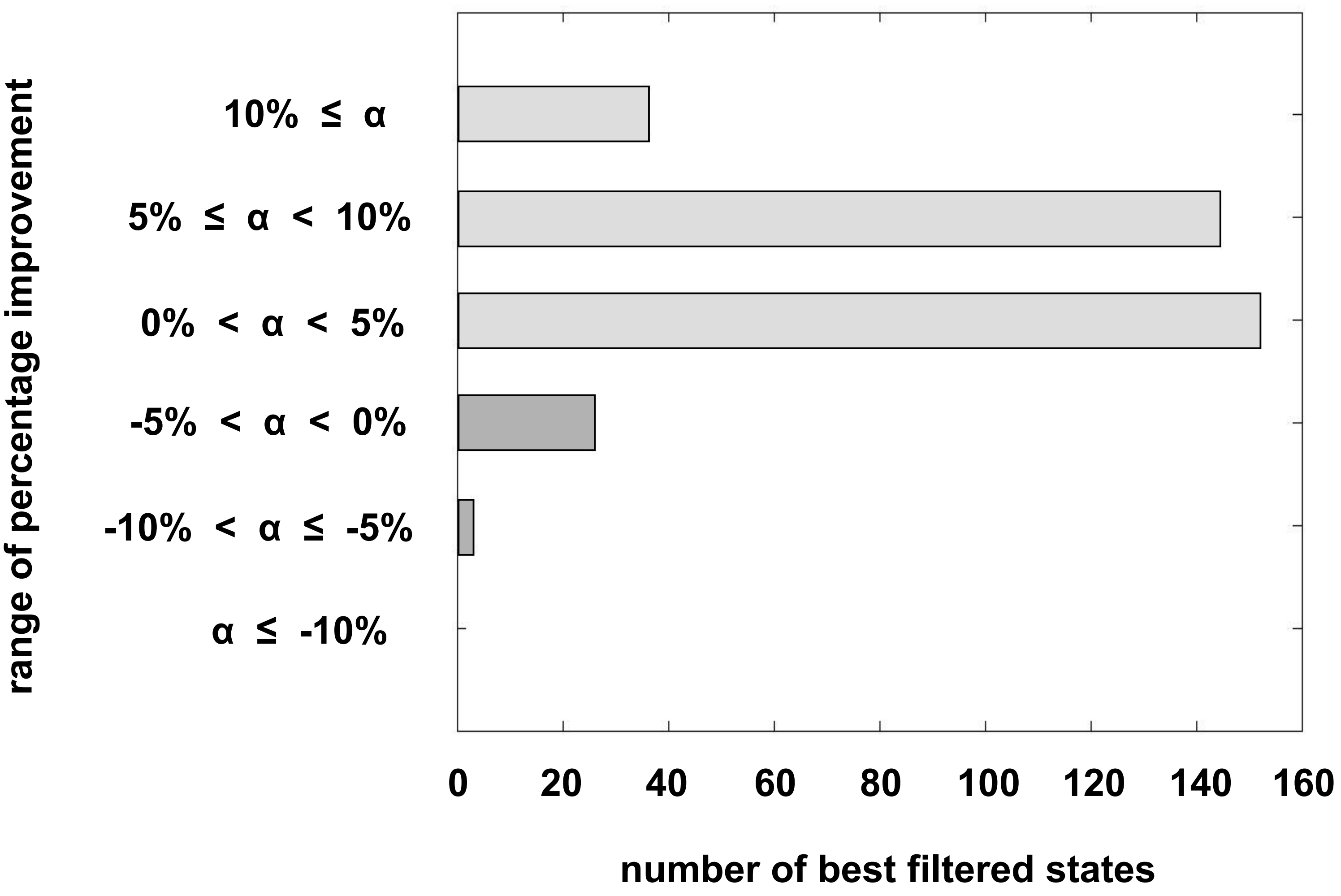
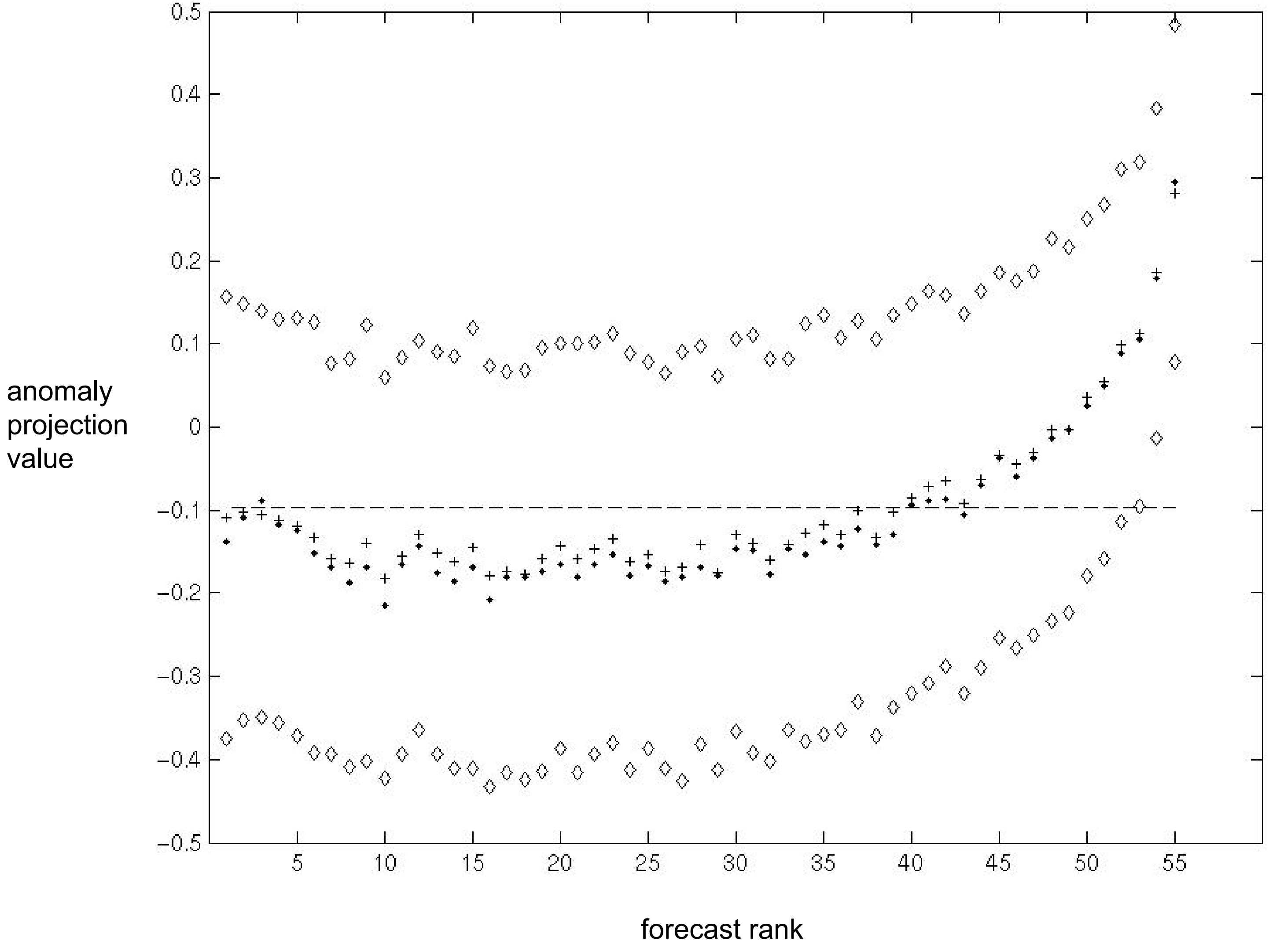


Figure 11. Mean and median anomaly projection value, and mean anomaly projection value plus or minus one standard deviation, as a function of forecast rank. Forecast rank is given on the abscissa and anomaly projection value is given on the ordinate. The "+" ("•") symbols indicate the mean (median) anomaly projection value associated with each forecast rank. The upper (lower) sequence of diamond symbols denotes for each forecast rank the mean anomaly projection value plus (minus) one standard deviation. The grey horizontal dashed line indicates the mean anomaly projection value for all forecast ranks.


anomaly projection value
forecast rank