Using Filtering to Mitigate Stochastic Model Errors' Effect on Ensemble Covariance. Part II: Employment of Filtered States in Hybrid Ensembles.

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

McLay and Martin (2005) introduce a post-processing method that employs filtering to mitigate the effect of stochastic model errors on ensemble covariance. They define a prototype filtering scheme and evaluate its efficacy through composite and ensemble-by-ensemble comparisons between the error characteristics of the filtered states and those of the operational ensemble. These comparisons suggest that the filtering scheme can consistently produce a set of states which are generally less corrupted by stochastic errors than the operational members are. The positive results of the comparisons serve as the impetus for investigating the use of the filtered states in hybrid ensembles.

In this paper prototype hybrid ensembles are defined and used as the basis for a two-part analysis of outstanding issues related to hybrid ensembles' variance, covariance, and associated probabilistic forecasts. First, the variance and covariance of the prototype hybrid ensembles are compared with the respective values for the operational ensemble through a set of idealized experiments in which the statistical distribution of the stochastic errors is assumed known. Second, the prototype hybrids are constructed for each operational ensemble in a diverse sample, and the performance of multi-dimensional probabilistic forecasts derived from the hybrids is systematically compared with that of similar forecasts derived from the operational ensemble. The results of the analysis suggest that some of the prototype hybrid ensembles are able to offer better covariance and improved probabilistic forecasts, and support further investigation of the methodology.

1. Introduction

McLay and Martin (2005) introduce a method of ensemble post-processing designed to counter 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 stochastic errors. It then uses some number of the filtered states to complement or supplant the operational members, forming a so-called hybrid ensemble. The hypothesis is that the composition of the hybrid ensemble will provide it with covariance that is improved relative to that of the operational ensemble. With improved covariance, and mean and variance comparable to that of the operational ensemble by design, the hybrid ensemble should yield multi-dimensional probabilistic forecasts that are better than those derived from the operational ensemble.

McLay and Martin (2005) commence scrutiny of the proposed methodology by testing a prototype filtering scheme that utilizes so-called pair-wise filtering: the formation of all possible pairs of operational members followed by the averaging of the members in each pair. They demonstrate, in particular, that 1) For any given ensemble forecast the pair-wise filtered states are, in general, less corrupted by stochastic errors than any operational members are, including the ensemble mean, and 2) A hybrid ensemble that is constructed using the pair-wise filtered states is unlikely to have a mean that is significantly different from that of the operational ensemble.

This study extends the scrutiny of the proposed methodology by considering several still outstanding issues related to the hybrid ensembles' variance, covariance, and associated probabilistic forecasts. These issues are, specifically: Can a hybrid ensemble that is constructed using filtered states have variance comparable to that of the operational ensemble? Is it, in fact, possible for certain versions of hybrid ensemble to offer covariance better than that of the operational ensemble? Assuming that it is possible, can these versions of hybrid ensemble offer improved probabilistic forecasts? Here, these issues are

addressed through a two-part investigation of prototype hybrid ensembles constructed using the pair-wise filtered states. First, the variance and covariance of the prototype hybrid ensembles are compared with the respective values for the operational ensemble through a set of idealized experiments in which the statistical distribution of the stochastic errors is assumed known. Second, the prototype hybrids are constructed for each operational ensemble in a diverse sample, and the performance of multi-dimensional probabilistic forecasts derived from the hybrids is systematically compared with that of similar forecasts derived from the operational ensemble.

2. Data and Verification Measures

a. Ensemble Data

Analysis is based upon 221 different National Center for Environmental Prediction (NCEP) Global Forecast System (GFS) 0000UTC initialization 11-member ensemble forecasts of 192h 500 hPa geopotential height. These ensembles were generated during the period between 21 December 2002 and 31 July 2003. The data were obtained on 2.5°-by-2.5° latitude-longitude grids in a cylindrical equidistant (CED) projection. The appropriate 0h leadtime control forecast was used as verification in all forecast error calculations. The 12 July 2003 ensemble is missing and hence unverifiable and, as a further consequence, the 192h leadtime ensemble forecasts initialized 4 July 2003 are unverifiable. The above sample of ensemble forecasts is modest in size but describes a relatively diverse array of flow types since it includes flows from the depths of winter to the peak of summer.

b. Climatological Data

All climatological 500 hPa geopotential height values are derived from NCEP reanalysis data. These data were obtained from the National Center for Atmospheric Research in the form of 2.5°-by-2.5° latitude-longitude grids for each day in the 30-year period 1972

to 2002. The data provide 30 different 500 hPa height values for each day and gridpoint, and these 30 different values are averaged to obtain the climatological value for the day and gridpoint.

c. Standard Verification Measures for Dichotomous Probability Forecasts

The Brier score is defined as

$$BS = \frac{1}{n} \sum_{k=1}^{n} (y_k - o_k)^2 \ ,$$

where n is the number of forecasts in the sample, y_k is the ensemble-derived event probability for a given forecast, and o_k is the observed event probability for a given forecast. The latter assumes a value of '1' if the event occurs and '0' if the event doesn't occur. Note that $0 \leq \mathrm{BS} \leq 1$, and that $\mathrm{BS} = 0$ for a perfect ensemble. The Brier score can be interpreted as being the "mean-squared error" of probability forecasts (Wilks 1995).

The relative operating characteristic (ROC) is based upon the 2x2 contingency table illustrated in Fig. 1. Using this table, one can define a hit rate as $H \,/\, (H + M)$ and a false alarm rate as $F \,/\, (F + R)$. These two quantities suffice to characterize an ensemble's performance for a given event probability threshold: The higher the hit rate and the lower the false alarm rate, the better the ensemble's performance. The ROC is an assessment of the ensemble's performance in terms of hit rate and false alarm rate for the entire range of probability thresholds. The assessment involves first calculating the hit rate and false alarm rate for each of the ensemble's allowable probability thresholds. Following this, a plot is made of hit rate versus false alarm rate, where the x-axis (y-axis) defines the false alarm rate (hit rate) and each point on the plot is determined by the hit rate and false alarm rate for a given probability threshold. Assuming that the ensemble has some positive skill, the plotted points will describe a curve that is arched toward the

upper left corner of the plot. The area under this curve is the measure of the level of skill, with an area of 1.0 being associated with a perfect ensemble. In practical terms, the ROC evaluates the ability of an ensemble to discriminate between those occasions when an event is likely and those when it is not likely.

3. Prototype Hybrid Ensembles

a. Configuration

The construction of a hybrid ensemble begins with a determination of whether to complement or supplant the operational members with some number of filtered states. This determination ascribes to the hybrid ensemble one of two configurations. In the case that the operational members are complemented with filtered states the hybrid ensemble assumes the configuration (hereafter referred to as Configuration A) given in Fig. 2a. All 11 operational members are used together with some arbitrary number of filtered states n_f. In the case that the operational members are supplanted with filtered states the hybrid ensemble assumes the configuration (hereafter referred to as Configuration B) given in Fig. 2b. Some numbers m and n_f of operational members and filtered states, respectively, are used, where n_f is arbitrary and in general $m < n_f$. Whether one configuration is more effective than the other is unknown. Configuration B may actually offer the best performance, since the proportion of total members that are reduced-error filtered states is likely to be greatest in it. However, Configuration A is arguably the simpler of the two since it is just built around the whole of the operational ensemble. On the basis of this simplicity alone, the present analysis is restricted to investigation of hybrid ensembles of Configuration A.

b. Filtered state Selection Procedure

Having settled upon a configuration for the hybrid ensembles, the remaining task is to define how to select the filtered states to be used in a given hybrid. Considerable heed

must be paid to the variance of the hybrid ensemble in any definition, as experiments (not shown here) reveal that a selection procedure that operates independently of this variance (e.g. one based upon random selection of filtered states) can easily culminate in a hybrid ensemble that is substantially less variant than the operational ensemble. Such a hybrid would be unacceptable given both the post-processing method's requirement that the hybrid have variance comparable to that of the operational ensemble and the general tendency for operational ensembles to be sub-variant to begin with. With consideration of variance given precedence, then, the following selection procedure is adopted for the present analysis. First, one of the 55 filtered states is individually added to the 11-member operational ensemble to form a 12-member hybrid ensemble. For this 12-member hybrid ensemble, some measure of the variance of the height values at the pair of gridpoints of interest is calculated. This process is repeated independently for all 55 filtered states. The 12-member hybrid ensemble, ${h_{12}}^{max}$, with the greatest measure of variance is identified, and the filtered state s_{12} associated with this hybrid ensemble is selected and set aside. Next, one of the 54 remaining filtered states is individually added to the 12-member hybrid ensemble ${h_{12}}^{max}$ to form a new 13-member ensemble. For this 13-member ensemble, some measure of the variance of the height values at the pair of gridpoints of interest is calculated. This process is repeated independently for all 54 of the remaining filtered states. The 13-member hybrid ensemble, ${h_{13}}^{max}$, with the greatest measure of variance is identified, and the filtered state s_{13} associated with this hybrid ensemble is selected and set aside with s_{12}. In the same way, hybrid ensembles ${h_{14}}^{max}$, ${h_{15}}^{max}$,...., ${h_n}^{max}$ are identified, and filtered states s_{14}, s_{15},...., s_n are set aside. The sequence is completed when the hybrid ensemble ${h_n}^{max}$ of some predetermined size n has been identified, where $n = 11 + n_f$. Recognize that each iteration within the procedure outlined above does only the following: It obtains the most variant hybrid ensemble ${h_i}^{max}$ that can be realized by adding some individual filtered state to a specific hybrid ensemble ${h_{i-1}}^{max}$. The iteration does not ensure that ${h_i}^{max}$ is the absolute most

variant hybrid ensemble of i members that can be found. That is, there may exist another i-member hybrid ensemble ${h_i}^*$, with a different composition of filtered states than ${h_i}^{max}$, that is more variant than ${h_i}^{max}$. Similarly, at the conclusion of all the iterations ${h_n}^{max}$ is not ensured to be the absolute most variant hybrid ensemble of n members. This circumstance could be regarded as a limitation of the above procedure. However, the procedure is attractive for its simplicity and computational efficiency.

4. Idealized Experiments with the Hybrid Ensembles

a. Methodology

A small set of idealized experiments is used to gain insight into whether the prototype hybrids defined in Section 3 actually can offer improved covariance, whether certain values of n_f are characteristically associated with any hybrids that do have improved covariance, and how the variance of these hybrids compares with that of the operational ensemble. The experiments are each based upon two-dimensional ensemble distributions derived from operational ensemble data for a pair of gridpoints (Fig. 3). The operational ensemble data that is used in all the experiments is taken from 100 different randomly selected operational ensembles. Each experiment involves completing the following so-called "core" procedure 100 times by using the data from each of the 100 different operational ensembles in succession.

The Core Procedure

1. Take a given operational ensemble. For any two gridpoints this operational ensemble provides 11 pairs of height values (each pair being comprised of a given ensemble member's values at the two gridpoints). Each pair of height values is perturbed by first randomly drawing an "error" vector from some pre-specified two-dimensional distribution with means μ_1 and μ_2 in the respective dimensions, variances σ_1^2 and σ_2^2, and correlation coefficient ρ, and then subtracting this error vector from the vector comprised of the

pair of height values. Since there are 11 pairs of height values, eleven error vectors are randomly drawn. The collection of eleven perturbed pairs of height values is taken to represent the sample that the operational ensemble would provide were it not affected by stochastic model errors. In other words, the collection of eleven perturbed pairs represents a sample from the true forecast distribution, assuming the operational ensemble suffers no bias.

2. A hybrid ensemble with n_f filtered states is constructed from the original, unperturbed operational ensemble.

3. The covariance of the hybrid ensemble (cov_h), original operational ensemble (cov_o), and perturbed operational ensemble (cov_p) are each calculated. The difference between cov_h and cov_p (d_{hp}) and between cov_o and cov_p (d_{op}) are calculated, and the magnitudes of the two differences are compared. Note is taken of the smaller of the two differences. Additionally, the variance of the hybrid ensemble is calculated at each of the two gridpoints, as is that of the original operational ensemble, and a comparison between the variance values of the two ensembles is subsequently made at each gridpoint.

4. Steps 1-3 are repeated a large number of times, say 1000, and the proportion of these 1000 trials in which d_{hp} is smaller than d_{op} is tabulated (this proportion is referred to as α).

5. Steps 1-4 are repeated for all possible combinations of values for σ_i^2 and n_f, where σ_i^2 is allowed to assume the values 10^2, 15^2, 20^2, 30^2, 40^2, and 50^2 m^2, and n_f is allowed to assume the values 5, 10, 15, 20, 25, and 30. At the conclusion of Step 5 a table of α as a joint function of σ_i^2 and n_f is obtained. An example of such a table is provided in Fig. 4.

Each $\alpha(n_f, \sigma_i^2)$ provides an indication of the likelihood that a hybrid ensemble with n_f filtered states will yield better covariance than the operational ensemble when the operational ensemble is arbitrarily corrupted by stochastically-induced errors with statistics

μ_i, σ_i^2, and ρ. The idea is that if the hybrid concept is to have value then in a variety of different error-statistic scenarios there must be certain hybrid ensembles that can provide α values greater than .5. For the purpose of the experiments in the current study, a hybrid that provides an α value greater than or equal to .55 is considered likely to yield better covariance for the given error statistics, and a hybrid that provides an α value less than or equal to .45 is considered unlikely to yield better covariance. A hybrid that provides an α value between .46 and .54 is considered to yield covariance that is indistinguishable from that of the operational ensemble.

To ensure that the experiments are of manageable scope, the same type of error distribution is used for the duration of each experiment, and only two types of error distribution are considered for the experiments: normal and uniform. Furthermore, the two-dimensional distributions are constrained to have the same mean and variance in each dimension (i.e. a given error distribution's mean is the same at each gridpoint, as is the distribution's variance). The mean of each error distribution is set to zero, since the interest of the study is with stochastic, and not systematic, error effects. Lastly, the same value of the error distribution's correlation coefficient ρ is used for the duration of each experiment. The particular value that the correlation coefficient is assigned for a given experiment is one of the seven values -6., -.4, -2., 0., .2, .4, and .6. These two simplifications mean that the total number of experiments completed is 14: seven experiments derive from using a uniform error distribution and each of the seven ρ values, and another seven derive from using a normal error distribution and each of the seven ρ values.

In a final effort to keep the experiments of manageable scope, all experiments are constrained to deal only with instances wherein at each gridpoint the operational ensemble's variance is less than that of the perturbed operational ensemble's variance (i.e. instances wherein the stochastic errors affecting the operational ensemble are such that the operational ensemble's variance at each gridpoint is less than what it would be if

the ensemble were not affected by these stochastic errors). Such instances are the focus given the general tendency for operational ensembles to be sub-variant. As a result of this focus, for the purpose of these experiments the measure of variance that is adopted for the hybrid ensemble construction procedure is defined to be *the average variance at the two gridpoints*. This measure is simple yet consistent with the fact that in the above scenario the construction procedure should operate to give the hybrid ensemble as large a variance value as possible at each gridpoint.

b. Observations

Observations from the experiment based upon the normal distribution with $\mu = 0$ and $\rho = +.4$ exemplify the experimental observations as a whole, and are highlighted here. To begin with, the observations indicate that for almost every one of the dates (95 of 100) at least one of the hybrids with n_f values of 5, 10,..., 30 is likely to have covariance that is better than that of the operational ensemble. Bolstering this indication is the fact that for each of the dates, on average, three of the hybrids with n_f values of 5, 10,..., 30 are likely to have covariance that is better. As illustration, Fig. 5 presents α as a function of n_f for the case where $\sigma^2 = 400\ m^2$ for two dates, one in the winter and one in the summer. The plot of Fig. 5a indicates that for 18 February 2003 hybrids with n_f = 5, 10, or 15 are likely to have covariance that is better than that of the operational ensemble, while the plot of Fig. 5b indicates that for 21 July 2003 hybrids with n_f = 5, 10, 15, or 20 are likely to have covariance that is better.

Observations from the experiment also suggest that even when one constructs hybrid ensembles from a relatively narrow range of n_f values (one much smaller than 5,10,...,30) one can still find for a large majority of dates hybrids that have covariance that is better than that of the operational ensemble. For example, for 71 of 100 dates one finds that one or the other of the hybrid ensembles with n_f values of 10, 15, or 20 is likely to have better covariance, and for 63 of 100 dates one finds that one or the other of the hybrid ensembles

with n_f values of 10 or 15 is likely to have better covariance. Additionally, observations suggest that one can even limit attention to those hybrid ensembles constructed from a specific n_f value and still be likely to obtain covariance that is better than that of the operational ensemble for a majority of the 100 dates. For example, hybrid ensembles with $n_f = 15$ are likely to have better covariance for 59 of 100 dates, and hybrid ensembles with $n_f = 5$ and $n_f = 20$ are likely to have better covariance for 56 of 100 dates.

The crux of the above results is that the prototype hybrid ensembles can indeed have better covariance than the operational ensemble does. In fact, for a large majority of dates there are usually multiple hybrid ensembles that are likely to exhibit better covariance. Also apparent from these results is that the hybrid ensembles with improved covariance generally can be constructed with only a small fraction of the total of 55 filtered states.

Another important observation from the experiment is that the values of n_f associated with hybrid ensembles that have improved covariance vary somewhat from date to date. For example, while for 63 of 100 dates one or the other of the hybrid ensembles with $n_f = 10$ and 15 is likely to provide better covariance, the hybrid ensemble with $n_f = 10$ is likely to provide better covariance for only 50 of the 100 dates, and the hybrid ensemble with $n_f = 15$ is likely to provide better covariance for 59 of the 100 dates. Furthermore, there are 37 dates for which hybrid ensembles with $n_f = 10$ or 15 are not likely to provide improved covariance. In view of this, it would be ideal to make the construction of hybrid ensembles flow-dependent, such that the appropriate value of n_f to use is determined date by date. As the above observations suggest, however, it would probably be sufficient to vary the values of n_f among a narrow group of choices, say among 10, 15, and 20, or among 10, 15, and 30.

The above observations are specifically from the experiment with the normal distribution with $\rho = +.4$, and $\sigma^2 = 400\ m^2$. They are however, highly indicative of results gained for the same experiment with σ^2 values greater than or equal to 225 m^2. The

hybrid ensembles are somewhat less effective in the case that $\sigma^2 = 100\ m^2$, but since this case reflects a scenario in which the stochastically-induced errors of the ensemble members are quite small it is not of great consequence.

The above results are also not unlike those obtained from experiments involving the normal distribution and values of ρ other than +.4. In fact, the contrasts among the experimental results are few and readily characterizable. First, the number of dates for which hybrids that have improved covariance can be found diminishes as the error distribution's correlation approaches zero (from either the negative or positive direction), but such hybrids can be found for a majority of dates even when the correlation is as small as ± 2. Second, hybrids that provide improved covariance can be found for a slightly greater proportion of the 100 dates if the correlation of the error distribution is positive than if it is negative. Finally, hybrids that provide improved covariance are least prevalent when the correlation of the error distribution is zero. For example, for the experiment with the normal distribution and $\rho = 0$ with $\sigma^2 = 400\ m^2$ such hybrids can be found for only 46 of the 100 dates. However, in the experiments with $\rho = 0$ it is very easy to find for almost every date hybrids whose covariance is indistinguishable from that of the operational ensemble. Thus, when the correlation of the error distribution is zero the hybrid ensembles usually have covariance that is as good as that of the operational ensemble, and a not insignificant proportion of the time some of the hybrid ensembles are likely to have covariance that is better.

A final discussion concerns how the variance of the hybrid ensembles compares with that of the operational ensemble. The results indicate that the variance of hybrids with $n_f \leq 15$ actually may exceed that of the operational ensemble. Specifically, for $n_f = 5$, 10, and 15 this is the case for 95, 61, and 8 of the 100 dates, respectively. It's also found that hybrids with n_f values as large as 25 will have variance that is $\geq 80\%$ of that of the operational ensemble on a regular basis. For instance, a hybrid with $n_f = 25$ (20) has variance $\geq 80\%$ of that of the operational ensemble for 63 (100) of the 100 dates.

The overarching implication is that hybrids with $n_f \leq 25$ will generally have variance that is fairly comparable to that of the operational ensemble, but that hybrids with $n_f \geq 15$ will usually involve some trade between reduced variance and any improvement in covariance.

As a last note, the experiments using the uniform distribution did not produce results notably different from those associated with the experiments using the normal distribution. Hence, in the interest of brevity the results related to the uniform distribution are not discussed.

5. Evaluation of the Hybrid Ensembles' Probabilistic Forecasts

a. Philosophy

The idealized experiments are illuminating, but they cannot completely characterize the hybrid ensembles' performance potential. For instance, the actual frequency at which the hybrid ensembles are found to have improved covariance is liable to differ somewhat from that seen in the idealized experiments, since the stochastic errors associated with real numerical weather prediction models might derive from statistical distributions other than those defined in the experiments. Also, it is unknown whether the improvement in covariance obtained via the hybrid ensembles is sufficient to translate into probabilistic forecasts that are improved relative to those from the operational ensemble. Similarly, it is unknown whether the trade between improved covariance and decreased variance that is associated with hybrid ensembles with relatively large values of n_f will prevent these hybrids from providing improved probabilistic forecasts. In view of these outstanding points there is a need to examine probabilistic forecasts derived from the hybrid ensembles. This section specifically presents an evaluation of probabilistic forecasts derived from the prototype hybrid ensembles described in Section 3. Simplicity is given precedence and attention is restricted to versions of these prototype hybrids in which the number of filtered states is constant from date to date. Since neither the number of

filtered states nor the construction procedure (i.e. the configuration and filtered state selection process) changes from date to date for each version of hybrid investigated in the present analysis, these versions can be thought of as being flow-independent.

Given the results of the idealized experiments, the flow-independent hybrids that are the focus of the present analysis might be expected to provide less-frequent improvement in covariance (and hence less-frequent improvement in probabilistic forecasts) as compared to flow-dependent hybrids. However, the results for the flow-independent hybrids will serve as a useful benchmark. Another point regarding the interpretation of the present analysis has to do with the particular forecast verification procedure that is adopted. This procedure entails the construction of the flow-independent hybrids for a diverse but solitary sample of forecasts, followed by the identification of those hybrids that systematically yield the best probabilistic forecasts over the sample. Since the analysis highlights the best performance over a single sample, the results of the analysis are to be considered best-case results for the flow-independent hybrids.

b. Specific Ensemble-derived Probability Forecasts to be Investigated

As in the idealized experiments, the work in this section is constrained to deal solely with two-dimensional ensemble distributions of 192h 500 hPa geopotential height. An example of such a distribution is shown in Fig. 6. The probabilities derived from these distributions must be defined for a *finite* set of mutually exclusive, collectively exhaustive (MECE)[2] outcomes. One way to obtain such a set of outcomes is as follows. First, a climatological height value is defined for each of the two gridpoints associated with the sample space. Once these two climatological values (c_1,c_2) are determined, the sample space is partitioned into four sample subspaces, as depicted in Fig. 7. Given these sample subspaces, a set of four MECE outcomes can be readily defined: outcome 1 is defined

[2]Outcomes are mutually exclusive if only one outcome can occur, and they are collectively exhaustive if they encompass the entire sample space as a set (Wilks 1995).

to be when the point corresponding to the verifying height values falls into S_1, outcome 2 is defined to be when the point corresponding to the verifying height values falls into S_2, and so forth. Not only is this set of MECE outcomes of manageable size, the set is also nominal, meaning that there is no inherent ordering of the outcomes based upon some measure of magnitude. For example, outcome 1 is not greater than outcome 2 in any measurable sense. The nominal nature of this set of outcomes means that forecasts for the set can be alternatively described in terms of forecasts for a sequence of four dichotomous events: event 1 is the occurrence (or, equivalently, nonoccurrence) of outcome 1, event 2 is the occurrence (or nonoccurrence) of outcome 2, and so forth. Probabilities for each of these events are forthcoming from a given ensemble joint distribution of height. For instance, if four of a total of 10 points in such a distribution lie in S_2, then the ensemble probability of occurrence of event 2 (i.e. the event that the point corresponding to verification ultimately lies in S_2) is 4/10, or 40%. Also, there are standard and relatively simple verification measures for probability forecasts of dichotomous events, and the probability forecasts for each of these events are verified independently of each other. Thus, for the current work verification of ensemble-derived probabilities is facilitated when the probabilities are formulated and dealt with in the context of the four dichotomous events described above.

Further consideration of the above four events reveals that they each associate a specific pair of signs with the pair of verifying height anomalies. For instance, if event 1 occurs, then the point corresponding to verification falls into S_1, and the verifying height anomalies include a negative (-) anomaly at gridpoint 1 and a negative (-) anomaly at gridpoint 2. Thus, event 1's occurrence associates a pair of negative signs, (-,-), with the pair of verifying height anomalies, where the first entry in the parentheses corresponds to the sign of the height anomaly at gridpoint 1, and the second entry corresponds to the sign of the height anomaly at gridpoint 2. To underscore the associations, Fig. 8 depicts the pair of height anomaly signs that will be associated with the verifying height

anomalies if the point corresponding to verification falls into any one of the sample subspaces $S_1, ..., S_4$. These associations give probabilistic forecasts of events 1-4 some practical meaning, since common patterns in the 500 hPa height are often characterized by their attendant patterns of height anomaly sign. As an example, the well known negative phase of the Pacific-North American (PNA) pattern (Fig. 9a) tends to be associated (in the month of January) with a (+,+) pair of height anomaly signs at the locations of the two gridpoints given in Fig. 9b. Thus, if a probabilistic forecast derived from a two-dimensional ensemble distribution of height for those gridpoints indicates that event 4 is likely, the forecast would support to a certain extent the occurrence of the negative phase of the January PNA pattern.

Ensemble joint height distributions for five different gridpoint pairs are assessed in the current work (Fig. 10). Each of the gridpoint pairs is selected such that the points comprising it are roughly superposed with locations of 500 hPa height anomaly maxima and/or minima in a recurring flow pattern. The two points of both pairs 1 and 2 are generally superposed with the PNA pattern's height anomaly maxima over the north Pacific Ocean and north Atlantic Ocean, while the two points of pair 3 are generally superposed with the PNA pattern's height anomaly maxima over the north Pacific Ocean and southeast United States (Fig. 11a). Pairs 1 and 2 were chosen to be fairly similar to enable a rudimentary assessment to be made of how sensitive the results of the current work are to gridpoint selection. The two points of pair 4 are generally superposed with the North Atlantic Oscillation (NAO) pattern's height anomaly maxima over western Greenland and the west-central Atlantic Ocean (Fig. 11b). The fifth and last gridpoint pair is positioned so as to span the continental United States, in consideration of the common forecast problem of whether the flow pattern over the United States will be characterized by a trough (possibly a negative height anomaly) in the west and a ridge (possibly a positive height anomaly) in the east, or vice-versa.

c. Measure of Variance

Recalling Section 3b, the construction of the prototype hybrid ensembles requires definition of a "measure of variance". For the two-dimensional hybrids of the current work, the measure must summarize in a suitable way the variance of the height values at a particular pair of gridpoints. There are at least four different definitions of the measure that might suffice. Define ${\sigma_1}^2$ to be the variance of Φ at gridpoint 1 and ${\sigma_2}^2$ to be the variance of Φ at gridpoint 2. Then the four definitions are:

measure 1: *The average of the variances at the two gridpoints:*

$$\overline{\sigma^2} = \left({\sigma_1}^2 + {\sigma_2}^2\right)/2\,.$$

measure 2: *Only the variance at gridpoint 1,* ${\sigma_1}^2$.

This measure might be opted for in the case that the variance at one of the two gridpoints is substantially less optimal/more deficient than is the variance at the other gridpoint.

measure 3: *Only the variance at gridpoint 2,* ${\sigma_2}^2$.

This measure is the counterpart to measure 2.

measure 4: *The average of the percentage relative improvement in variance at the two gridpoints.*

Define imp_1 to be the % improvement at gridpoint 1 of the variance of h_i relative to that of ${h_{i-1}}^{max}$. Specifically,

$$imp_1 \quad = \quad \frac{\sigma^2_{1(i)} - \sigma^{2(max)}_{1(i-1)}}{\sigma^{2(max)}_{1(i-1)}} \cdot 100.$$

Analogously, define imp_2 to be the % improvement at gridpoint 2 of the variance of h_i relative to that of ${h_{i-1}}^{max}$:

$$imp_2 \quad = \quad \frac{\sigma^2_{2(i)} - \sigma^{2(max)}_{2(i-1)}}{\sigma^{2(max)}_{2(i-1)}} \cdot 100.$$

Measure 4 is then defined as $\overline{imp} = (imp_1 + imp_2)/2$. Note that measure 4, unlike the previous three measures, represents a relative comparison between the variances in the new hybrid h_i and the variances in the old hybrid ${h_{i-1}}^{max}$. Note also that imp_1, imp_2, and $\overline{imp}$ may assume either positive or negative values.

There are four companion measures to variance measures 1-4. These arise as a result of there being more than one way of defining the mean to be used in each variance calculation. That is, one can choose to calculate the variance about a mean that is fixed for the duration of the selection process (specifically, the mean of the operational ensemble), or one can choose to calculate the variance about a mean that is updated with each selection of a filtered state.

d. Hybrid Ensemble Naming Convention

Hereafter, each version of hybrid ensemble is referred to with a designation of the following form:

$$\text{A.x.yy}^{(\prime)} \quad ,$$

where 'A' indicates that the hybrid is of Configuration A, 'x' indicates the variance measure used, and 'yy' indicates the number of filtered states selected. A 'prime' appended to the end of the designation indicates that the mean used in each variance calculation is updated (rather than fixed) during the filtered state selection process.

e. Sequence of Operations and Tabulated Results

Table 1 summarizes the specific sequence of operations in the forecast verification procedure, given the information in Sections 5a-d.

Table 2 gives the best-case Brier score results arranged by gridpoint pair and binary event. In a given box of the table, the column of three entries identifies the version of hybrid ensemble that yields the maximum relative improvement in BS for the given

gridpoint pair and binary event, and provides specific information about the nature of this relative improvement. Details on the meaning of each of the three entries are provided in Table 3. If there are two columns of entries in a given box, the maximum relative improvement in BS is attained using the hybrid version identified in the first column in the box. The purpose of the second column of entries will be explained later. An example shows how Table 2 is to be interpreted. First, consider the row of boxes corresponding to gridpoint pair 5 and, in particular, the box corresponding to event (+,-). It is found that hybrid version A.2.17 yields a relative improvement in BS of $\approx 6\%$ for probability forecasts related to event (+,-), and that this version yields some form of positive relative improvement for all four binary events. These results can be arranged in an alternative, graphical format as given in Fig. 12. In this figure, the percentage relative improvement in BS derived from version A.2.17 for each binary event is superposed on the sample subspace corresponding to each event. Summarizing from Fig. 12, version A.2.17 yields noteworthy relative improvement in BS for the event (+,-) simultaneous with lesser relative improvement in BS for the other three events. From a practical standpoint, since event (+,-) is consistent at some basic level with a flow pattern defined by a ridge over the western United States and a trough over the eastern United States, the implication from Fig. 12 is that the use of version A.2.17 may have improved probability forecasts of the occurrence (or non-occurrence) of such a pattern over the course of the verification sample. Rigorous substantiation of this suggestion is, however, beyond the scope of the present analysis.

Table 2 also serves to provide information regarding a hybrid version's ROC verification results. It does so on the basis that during the course of the analysis the ROC results were usually commensurate with the BS results: Improvements (relative to the operational ensemble) in the BS of around 2% or less meant little if any improvement, but no degradation, in the ROC, while improvements in the BS of more than 2% generally, but not always, meant some modest improvement in the ROC. Also, the ROC was never

found to be improved when the BS was degraded. Figure 13 presents several examples which illustrate this guideline. First, consider Fig. 13a, which shows the ROC curves of hybrid version A.2.21 and the operational ensemble for gridpoint pair 1, event (-,+). For reference, hybrid A.2.21 yields $\approx 8\%$ relative improvement in BS for this particular gridpoint pair and event. It is easily seen in Fig. 13a that the area under hybrid A.2.21's ROC curve is somewhat greater than that under the operational ensemble's ROC curve. Thus, along with the noteworthy improvement in the BS, hybrid A.2.21 affords a modest but unambiguous increase in performance in terms of the ROC. Consider next Fig. 13b, which shows the ROC curves of hybrid version A.1.20 and the operational ensemble for gridpoint pair 4, event (-,+). For reference, hybrid A.1.20 yields $\approx 3\%$ relative improvement in BS for this gridpoint pair and event. As in Fig. 13a, it is fairly obvious in Fig. 13b that the area under the hybrid's ROC curve is somewhat greater than that under the operational ensemble's ROC curve. Thus, although hybrid A.1.20 yields a relatively minor improvement in the BS for this gridpoint pair and event, it still affords a modest but unambiguous increase in performance in terms of the ROC. As a last example, consider Fig. 13c, which shows the ROC curves of hybrid version A.1.19 and the operational ensemble for gridpoint pair 3, event (+,+). In this case, hybrid A.1.19 yields $\approx 1\%$ relative improvement in BS. The general indication of Fig. 13c is that the ROC curves of the hybrid and the operational ensembles are very similar. However, it can also be discerned that the area under the hybrid's ROC curve is slightly greater than that under the operational ensemble's curve. Therefore, while hybrid A.1.19 only delivers a very small improvement in the BS, it still manages to provide a commensurate improvement in the ROC curve.

Those few instances in Table 2 in which the ROC results do not adhere to the guideline and examples above are identified by there being two columns of entries in a given box. In these instances, the hybrid version identified in the first column in the box does not unambiguously yield relative improvement in the ROC, regardless of what the version

yields in terms of relative improvement in BS. The hybrid version identified in the second column in the box does give relative improvement in the ROC simultaneous with relative improvement in the BS, although the improvement in BS is less than that given by the hybrid version identified in the first column.

f. Synopsis of Results

The information in Table 2 can be summarized with regard to three questions of particular interest: 1) What is the magnitude of improvement in BS (ROC) for the binary events?, 2) Are the BS's (ROC's) for multiple binary events improved simultaneously?, and 3) The union of questions 1 and 2: Are the BS's (ROC's) for multiple binary events improved both notably and simultaneously?

Regarding element 1, assessment of the table indicates that for all gridpoint pairs a hybrid version can be found such that the BS for some binary event is improved by $\geq \approx$ 5% and such that the corresponding ROC curve is clearly improved. For 3 of 5 gridpoint pairs (1, 2, 4) a hybrid version can be found such that the BS for some binary event is improved by $\geq \approx 7\%$, and such that the corresponding ROC curve is improved. Also, for 3 of 5 pairs (1, 3, 5) hybrid versions can be found such that the BS's for two binary events are improved by $\approx 5\%$ (the improvement not necessarily being simultaneous), and such that the corresponding ROC curves are improved.

Regarding element 2, for 4 of 5 pairs (1, 2, 3, 5) a hybrid version can be found such that the BS's for three or more binary events are improved simultaneously. For 2 of 5 pairs (1, 5) a hybrid version can be found such that the BS's for all four binary events are improved simultaneously. The corresponding ROC measures proved less amenable to simultaneous improvement, because small improvements in BS were not necessarily accompanied by any definitive improvement in the ROC curves. Thus, usually (but not always) the number of binary events for which modest to notable improvement in ROC curves could simultaneously be attained was one less than the number of binary events

for which improvement in BS's could simultaneously be attained.

Regarding element 3, for 3 of 5 pairs (1, 2, 5) a hybrid version can be found that provides notable improvement (i.e. by $\geq \approx 6\%$) in the BS of one binary event simultaneous with some form of improvement in the BS of two or more other binary events. For 2 of these particular 3 pairs (1, 5) a hybrid version provides improvement in the BS of all four events. For 1 of 5 pairs (1) a hybrid version provides notable improvement in the BS of two binary events (8% for one event, 6% for the other) simultaneous with some form of improvement in the other two binary events. In these instances, the ROC measure is clearly improved for those binary events for which the BS is notably improved, and the ROC is improved modestly to marginally for those events for which the BS exhibits lesser improvement.

One additional observation deserves mention, this being that the prototype hybrids investigated here exhibit performance that is a fairly well behaved function of the number of filtered states admitted, n_f. In particular, for a given variance measure, analysis of the BS as a function of n_f reveals that there is usually one absolute minimum of BS, and that the BS values vary smoothly when the absolute minimum is approached either from the direction of increasing n_f or from the direction of decreasing n_f. Also, if any improvement can be attained with a given variance measure, then some form of improvement is generally attained independent of n_f so long as n_f does not become too large. Again, however, the most significant improvement in BS (ROC) is generally realized only with a specific number of filtered states admitted.

Collectively, the verification results suggest that modest, but systematic, performance gains in probabilistic forecasts of specific binary events are achievable through the use of hybrid ensembles. The results further suggest that systematic performance gains in forecasts of each of the four binary events are achievable simultaneously in some cases. In such cases there is also the suggestion that the performance gains associated with one or two of the binary events can be notable. These suggestions support the notion that

use of the hybrid ensemble concept can improve some probability forecasts.

4.4 Discussion and Conclusions

The objective of this study is to assess some outstanding questions regarding the use of filtered states in multi-dimensional hybrid ensemble forecasts. In the assessment, prototype hybrids are defined and subjected to a two-part analysis. First, the covariance and variance of the prototype hybrid ensembles are compared with the respective values for the operational ensemble through a set of idealized experiments in which the statistical distribution of the stochastic errors affecting the operational ensemble is assumed known. Specifically, for each experiment stochastic errors drawn from a normal or uniform distribution with zero mean and with prescribed values of variance and correlation coefficient are assumed to corrupt the operational ensemble. Each experiment focuses on two-dimensional ensemble and error distributions, and also focuses on the case wherein the stochastic errors cause the operational ensemble to have variance in each dimension that is less than what it would be in the absence of the errors.

The results of the idealized experiments support the idea that the hybrid ensembles can offer better covariance than the operational ensemble does. For instance, for a majority of dates in most experiments multiple hybrid ensembles can be found that are likely to exhibit better covariance. The results further suggest that the hybrid ensembles can be constructed so that they have improved covariance without an undue sacrifice in variance being incurred. In fact, indications are that the variance of hybrids constructed using 15 or fewer of the filtered states has some chance of exceeding that of the operational ensemble. The other main result of the idealized experiments is that the range of values of n_f (the number of filtered states used in constructing a hybrid) associated with hybrids that are likely to have improved covariance displays modest day-to-day variation. This means that it would be ideal to make the construction of hybrid ensembles flow-dependent, such that the appropriate number of filtered states to use is determined

date by date.

The second part of the analysis involves assessing whether hybrid ensembles are capable of yielding multi-dimensional probabilistic forecasts that perform better on a systematic basis than corresponding forecasts yielded by the operational ensemble. In the assessment, various flow-independent versions of the prototype hybrid ensembles are constructed for each of 221 different 11-member NCEP GFS 0000UTC F192 500 hPa geopotential height ensembles that span the period 0000UTC 21 December 2002 and 0000UTC 31 July 2003. Each of these hybrids is a two-dimensional distribution of geopotential height for a given pair of gridpoints. These hybrids are used to derive probabilistic forecasts for four dichotomous events that describe the specific signs of the verifying geopotential height anomalies at the pair of gridpoints. The performance of the forecasts is evaluated using two standard dichotomous probability forecast verification measures, the Brier score (BS) and the relative operating characteristic (ROC). The best results from the performance evaluation are compared to the results of a similar evaluation undertaken for forecasts derived from the operational 11-member ensemble. The hybrid ensembles are constructed and their associated probabilistic forecasts are evaluated for a total of five different gridpoint pairs.

The findings associated with the BS and ROC standard verification measures suggest that improvement in multi-dimensional probabilistic forecasts is in fact achievable through the use of filtered states in a hybrid configuration. For instance, for a majority of pairs (3 of 5) a version of hybrid can be found that provides notable (i.e. $\geq 6\%$) improvement in the BS of one event simultaneous with some form of improvement in the BS's of two or more other events. Also, for some of the pairs (2 of 5), a version of hybrid provides notable improvement in the BS of one event simultaneous with some form of improvement in the BS's of all three other binary events. For one of the five pairs, a version of hybrid is found that provides notable improvement in the BS's of two events (the improvement for one being $\approx 8\%$ and the other $\approx 6\%$), simultaneous with some

form of improvement in the BS's of the other two events. In the above instances, the hybrid ensembles generally provide improvement in the ROC measure that is commensurate with that in the BS. It is entirely possible that the hybrid versions investigated here yield larger gains in performance in association with other gridpoint pairs, as the choice of gridpoint pairs does not reflect any attempt to isolate those pairs associated with the best results.

The above results are most meaningful when it is recalled that they were obtained by testing flow-independent versions of the simplest of two hybrid configurations. Just as important, the testing measured performance over a sample of long-leadtime (192h) forecasts that spans a broad range of flow regimes and multiple seasons. When viewed in the context of these two facts, the results serve as strong suggestion that more advanced implementations of the hybrid ensemble concept might yield comparable or even better results over more general forecast samples, and hence prove a worthwhile research endeavor.

In view of the above observations, the results presented here are fairly encouraging. However, the reality is that the results in this paper and in McLay and Martin (2005) derive from one prototype filtering scheme and one prototype hybrid ensemble configuration, and much work remains to be done on the proposed method. The problem of filtering needs to be revisited, because while the pair-wise scheme appears to provide adequate filtering of 192h 500 hPa geopotential height, it is unknown whether this scheme consistently provides an optimum level of filtering. Furthermore, it is unknown whether this scheme would provide optimum or even adequate filtering when applied at other leadtimes and to other atmospheric variables. The problem of hybrid ensemble configuration also needs to be revisited, as configuration B hybrid ensembles are not investigated in the current series of papers. Configuration B hybrids arguably offer the most promise of improved covariance, and need to be subjected to the idealized experimentation and forecast verification process described in this paper. One of the observations from the

idealized experiments in this paper requires additional investigation as well, this observation being that the range of values of n_f associated with hybrids that are likely to have improved covariance exhibits some day-to-day variation. Greater knowledge of this variation may allow the probabilistic forecast performance of configuration A hybrids to be substantially improved.

It is also imperative that the post-processing method be applied to multi-model ensemble data, and ensemble data comprised of a larger number of members than just eleven. Doing so will serve to evaluate the method's utility in operational settings of the near future, wherein ensemble datasets formed of a large number of members from multiple numerical models will be the norm. Application of the method should also be extended to other atmospheric fields, such as 850 hPa temperature, and to other forecast lead times, such as 240h (10 days) and 336h (14 days). Furthermore, the ability of the hybrid ensembles to serve as the basis for types of probabilistic forecasts other than those visited in this dissertation should be assessed. Lastly, the probabilistic forecasts derived from the hybrid ensembles need to be evaluated using other verification measures in addition to the BS and ROC. Carrying out all of this analysis will help define the limits of the method's effectiveness, which at this point are unknown.

Acknowledgments

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

References

McLay, J. G., and J. E. Martin, 2005: Using filtering to mitigate stochastic model errors' effect on ensemble covariance. Part I: Evaluation of a prototype filtering scheme. *Mon. Wea. Rev.*, submitted.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

Table Captions

Table 1. Sequence of operations in the probabilistic forecast verification procedure.

Table 2. Best-case Brier score results arranged by gridpoint pair and event. See text and Table 3 for interpretation.

Table 3. Interpretation of the column of three entries in a given box of Table 2.

Table 1. Sequence of operations in the probabilistic forecast verification procedure.

1.	For a given operational ensemble, create various versions of hybrid ensemble (A.1.5, A.1.10, A.1.15,…., A.2.5, A.2.10,…., A.4.5, A.4.10,…. etc.).
2.	Derive probabilistic forecasts for all four dichotomous events from the operational ensemble and each of the hybrid ensembles.
3.	Perform steps 1 and 2 for 221 operational ensembles.
4.	Use the Brier score and relative operating characteristic to verify the probabilistic forecasts, then look for the version of hybrid ensemble that provides the forecasts that are most improved relative to the operational forecasts.
5.	Do all the above for joint height distributions associated with five different gridpoint pairs.

Table 2. Best-case Brier Score results arranged by gridpoint pair and event. See text and Table 3 for interpretation.

Gridpoint Pair	Event (-,-)	Event (+,-)	Event (-,+)	Event (+,+)
1	2.19 ✓ 4	2.30 ✓ 4	2.21 8 4	2.21 6 4
2	1.5' ✓ 1 ; 4.5' ✓ 3	4.8' ✓ 3	4.15' 8 3	——
3	3.5 ✓ 1 ; 1.19 ✓ 3	1.12 5 2	1.20 ✓ 2	4.32 5 2
4	3.6 8 1 ; 3.15 7 2	——	4.30 ✓ 2	——
5	2.8' ✓ 3	2.17 6 4	2.5 6 4	3.10' ✓ 3

Table 3. Interpretation of the column of three entries in a given box of Table 2.

Entry 1	designation of the hybrid version
Entry 2	the percentage relative improvement attained – values are rounded to the nearest integer – positive values greater than 5 are highlighted in light grey. – positive values less than 5 are represented with a check mark
Entry 3	the number of events for the given gridpoint pair for which the hybrid version specified in Entry 1 is able to simultaneously yield some form of positive relative improvement

Figure Captions

Figure 1. Relative operating characteristic 2x2 contingency table. The four possible contingencies are a "hit", "miss", "false alarm", and "correct rejection". As an example of the table's interpretation, if an event is forecast and is ultimately observed, then a "hit" is recorded. An event is considered to be forecast if the ensemble-derived probability for the event exceeds a pre-specified threshold value. The allowable threshold values are a function of the number of members in a given ensemble. For instance, if an ensemble has three members, then the allowable threshold values are 0.0, .33, .67, and 1.0.

Figure 2. Illustration of possible hybrid ensemble configurations. The area within each panel describes a two-dimensional sample space. Operational ensemble members and pair-wise samples are represented as dots and X's, respectively. a) Configuration A. b) Configuration B. See the text for further interpretation of these configurations.

Figure 3. Orientation of the gridpoint pair referenced in the idealized experiments. The horizontal position of a particular gridpoint is indicated by a black dot.

Figure 4. Example of a table that provides α as a function of error variance and n_f. Values of error variance (m^2) are given along the top of the table, and values of n_f are given along the left-hand side.

Figure 5. Two examples of α as a function of n_f from the experiment where the errors are drawn from a normal distribution with $\mu = 0$ m, $\sigma^2 = 400$ m^2, and $\rho = .4$. a) Results obtained using the 18 February 2003 ensemble data. Values of n_f are given along the bottom of the plot. A bar describes the value of α for a given value of n_f according to the scale along the left-hand side of the plot. A dotted line denotes the height of a bar associated with $\alpha = .5$. b) Results obtained using the 21 July 2003 ensemble data. Otherwise same as for a).

Figure 6. Ensemble joint distribution of 500 hPa geopotential height at two gridpoints. Possible values of height Φ_1 (m) at a gridpoint located at 170.0W longitude, 52.5N latitude define the abscissa, and possible values of height Φ_2 (m) at a gridpoint located at 22.5W longitude, 60.0N latitude define the ordinate. Each point on the plot depicts a particular ensemble member's specific values of height at the two gridpoints. The ensemble members are associated with the 14 May 2003 NCEP GFS 192h ensemble. The grey point corresponds to the control ensemble member, and the black points correspond to the perturbed ensemble members.

Figure 7. As for Figure 6, except that the dotted lines identify the climatological values of 500 hPa geopotential height c_1 and c_2 at gridpoints 1 and 2, respectively, for 14 May 2003. Also, the labels S_1, S_2, S_3, and S_4 indicate the sample subspaces described in the text.

Figure 8. As for Figure 7, except that the pair of signs in each parentheses are the signs of the verifying 500 hPa geopotential height anomalies at gridpoints 1 and 2 in the case that the point corresponding to verification falls into a given sample subspace. The first entry in parentheses corresponds to the sign of the height anomaly at gridpoint 1, and the second entry corresponds to the sign of the height anomaly at gridpoint 2.

Figure 9. a) Composite 500 hPa geopotential height (m) anomaly pattern associated with the negative phase of the Pacific North American (PNA) pattern. Composite based upon NCEP/NCAR Reanalysis data for the month of January and the years 1968, 1969, 1971, 1972, 1982, and 1989. Contours every 15m. Negative height anomalies ≤ -15m shaded. Figure created using NOAA-CIRES/Climate Diagnostics Center interactive website. b) Horizontal positions of two particular gridpoints located at the 500 hPa vertical level. The position of each gridpoint is indicated by a black dot. The designation of each gridpoint is indicated by the number below the dot

Figure 10. Orientations of the five different gridpoint pairs referenced in the analysis. The horizontal position of a particular gridpoint is indicated by a black dot, and the two gridpoints in a given pair are connected by a black dotted line. The designation of a given gridpoint pair is indicated by the number alongside the black dotted line associated with the given pair.

Figure 11. a) Composite 500 hPa geopotential height (m) anomaly field associated with the positive phase of the Pacific North American (PNA) pattern. Composite based upon

NCEP/NCAR Reanalysis data for the month of January and the years 1977, 1981, 1983, 1985, 1988, and 1992. Contours every 15m. Negative height anomalies ≤ -15m shaded. b) Composite 500 hPa geopotential height (m) anomaly field associated with the positive phase of the North Atlantic Oscillation (NAO) pattern. Composite based upon NCEP/NCAR Reanalysis data for the month of January and the years 1984, 1986, 1987, 1988, 1989, 1990, 1991, 1993, and 1994. Contours and shading same as for a). Figures created using NOAA-CIRES/Climate Diagnostics Center interactive website.

Figure 12. Percentage relative improvement in Brier score derived from hybrid version A.2.17 for gridpoint pair 5 for each of the four dichotomous events (-,-), (+,-), (-,+), and (+,+). The four boxes represent the partitioned sample space (refer to Fig. 8), and the relative improvement is superposed on the sample subspace corresponding to each event. Positive relative improvement less than 5.0% is represented with a check mark.

Figure 13. ROC curves of the operational ensemble and a given hybrid ensemble for a particular gridpoint pair and event. The solid (dashed) line is the ROC curve of the operational (given hybrid) ensemble. Each open circle (dot) defines the operational (given hybrid) ensemble's false alarm rate and hit rate for a specific probability threshold. For reference, the dotted line is the ROC curve of an ensemble with zero skill.

a) ROC curves for the operational ensemble and hybrid ensemble version A.2.21, gridpoint pair 1, event (-,+).

b) ROC curves for the operational ensemble and hybrid ensemble version A.1.20, gridpoint pair 3, event (-,+).

c) ROC curves for the operational ensemble and hybrid ensemble version A.1.19, gridpoint pair 4, event (+,+).

Figure 1. Relative operating characteristic 2x2 contingency table. The four possible contingencies are a "hit", "miss", "false alarm", and "correct rejection". As an example of the table's interpretation, if an event is forecast and is ultimately observed, then a "hit" is recorded. An event is considered to be forecast if the ensemble-derived probability for the event exceeds a pre-specified threshold value. The allowable threshold values are a function of the number of members in a given ensemble. For instance, if an ensemble has three members, then the allowable threshold values are 0.0, .33, .67, and 1.0.

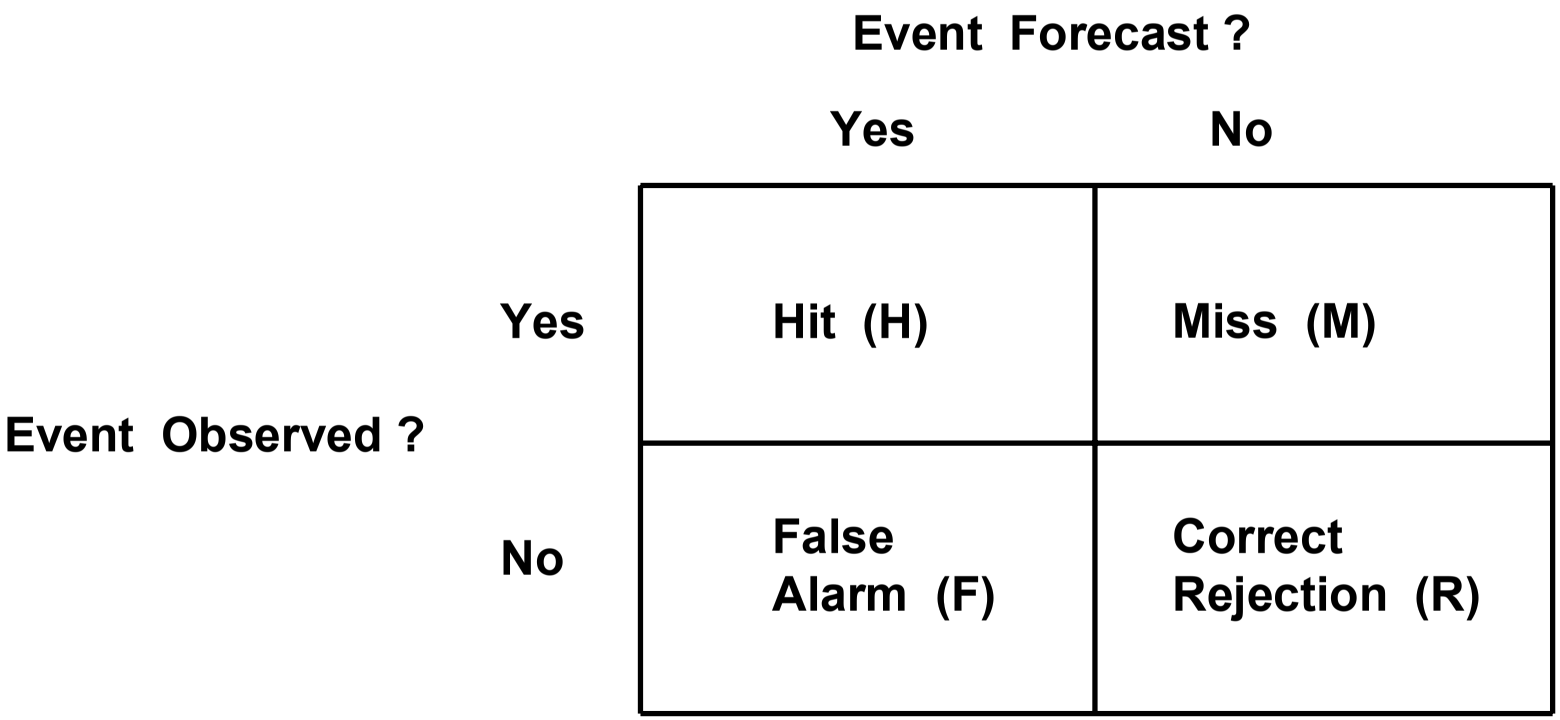

Figure 2. Illustration of possible hybrid ensemble configurations. The area within each panel describes a two-dimensional sample space. Operational ensemble members and pair-wise samples are represented as dots and X's, respectively. a) Configuration A. b) Configuration B. See the text for further interpretation of these configurations.

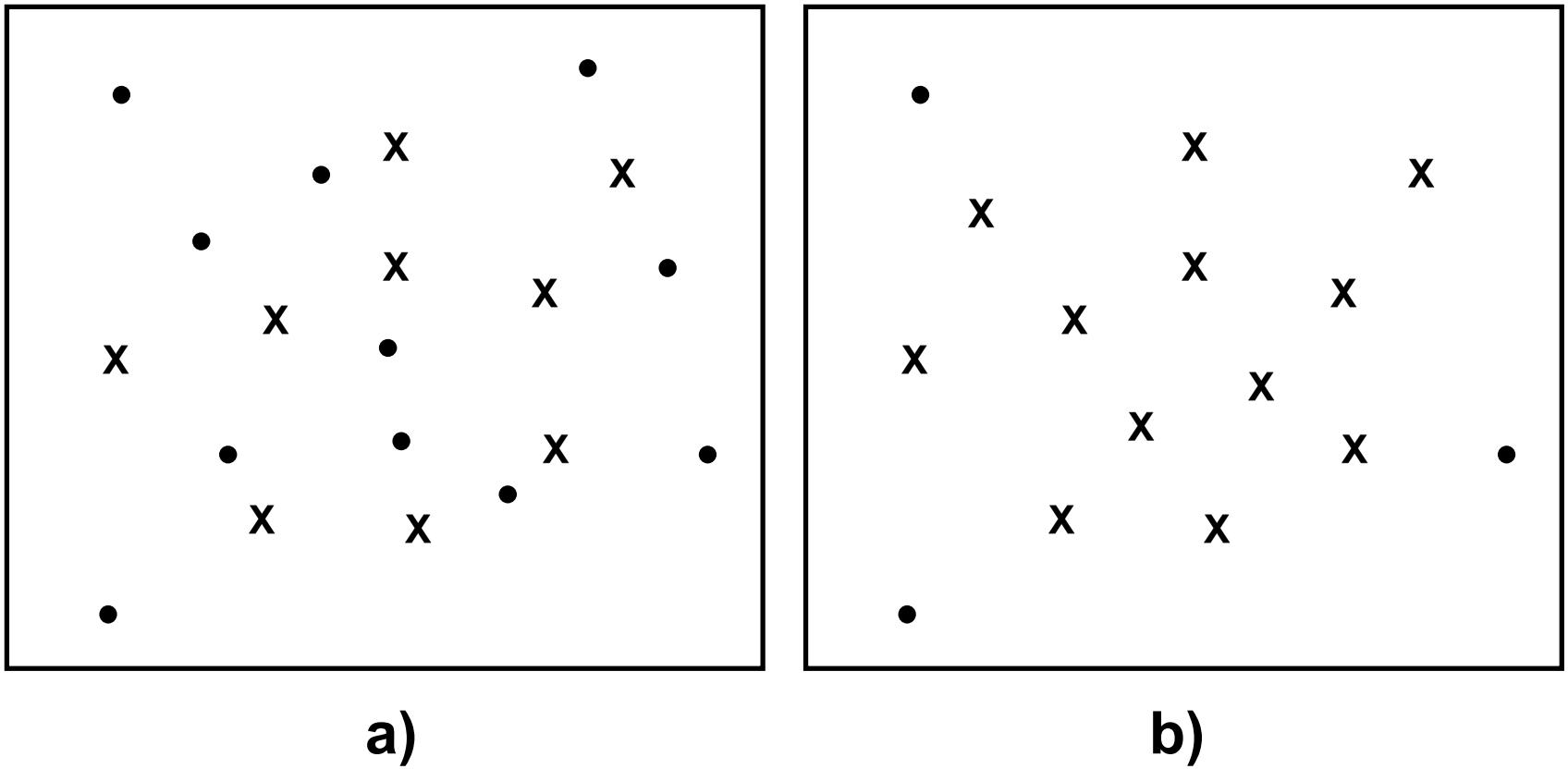

Figure 3. Orientation of the gridpoint pair referenced in the idealized experiments. The horizontal position of a particular gridpoint is indicated by a black dot.

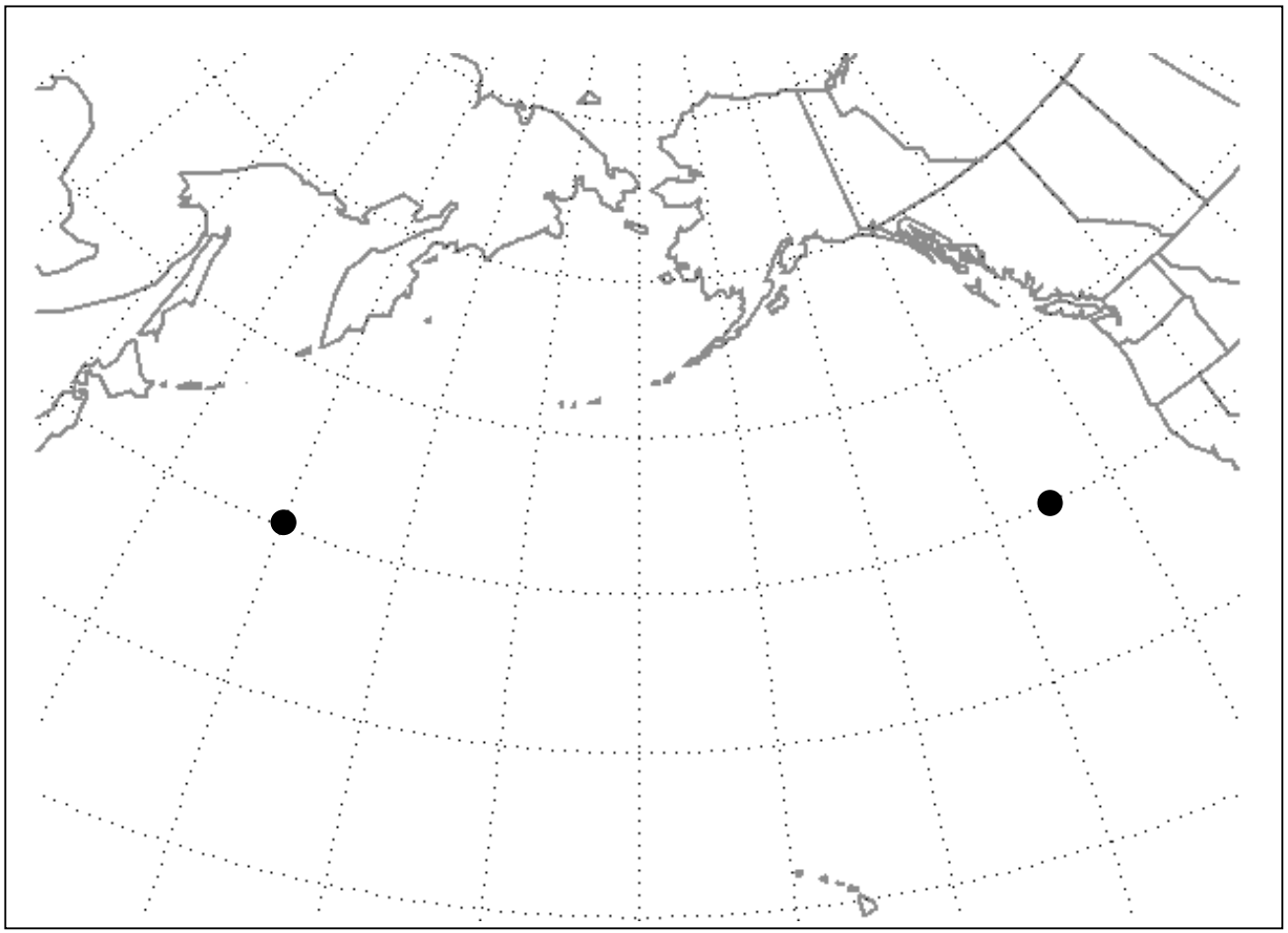

Figure 4. Example of a table that provides α as a function of error variance and n_f. Values of error variance (m^2) are given along the top of the table, and values of n_f are given along the left-hand side.

n_f \ σ^2	100	225	400	900	1600	2500
5	0.26	0.24	0.27	0.24	0.25	0.20
10	0.51	0.60	0.65	0.69	0.70	0.76
15	0.39	0.51	0.58	0.68	0.70	0.73
20	0.59	0.65	0.69	0.71	0.77	0.77
25	0.26	0.25	0.25	0.24	0.22	0.21
30	0.20	0.22	0.22	0.21	0.21	0.21

Figure 5. Two examples of α as a function of n_f from the experiment where the errors are drawn from a normal distribution with $\mu = 0$ m, $\sigma^2 = 400$ m^2, and $\rho = .4$. a) Results obtained using the 18 February 2003 ensemble data. Values of n_f are given along the bottom of the plot. A bar describes the value of α for a given value of n_f according to the scale along the left-hand side of the plot. A dotted line denotes the height of a bar associated with $\alpha = .5$. b) Results obtained using the 21 July 2003 ensemble data. Otherwise same as for a).

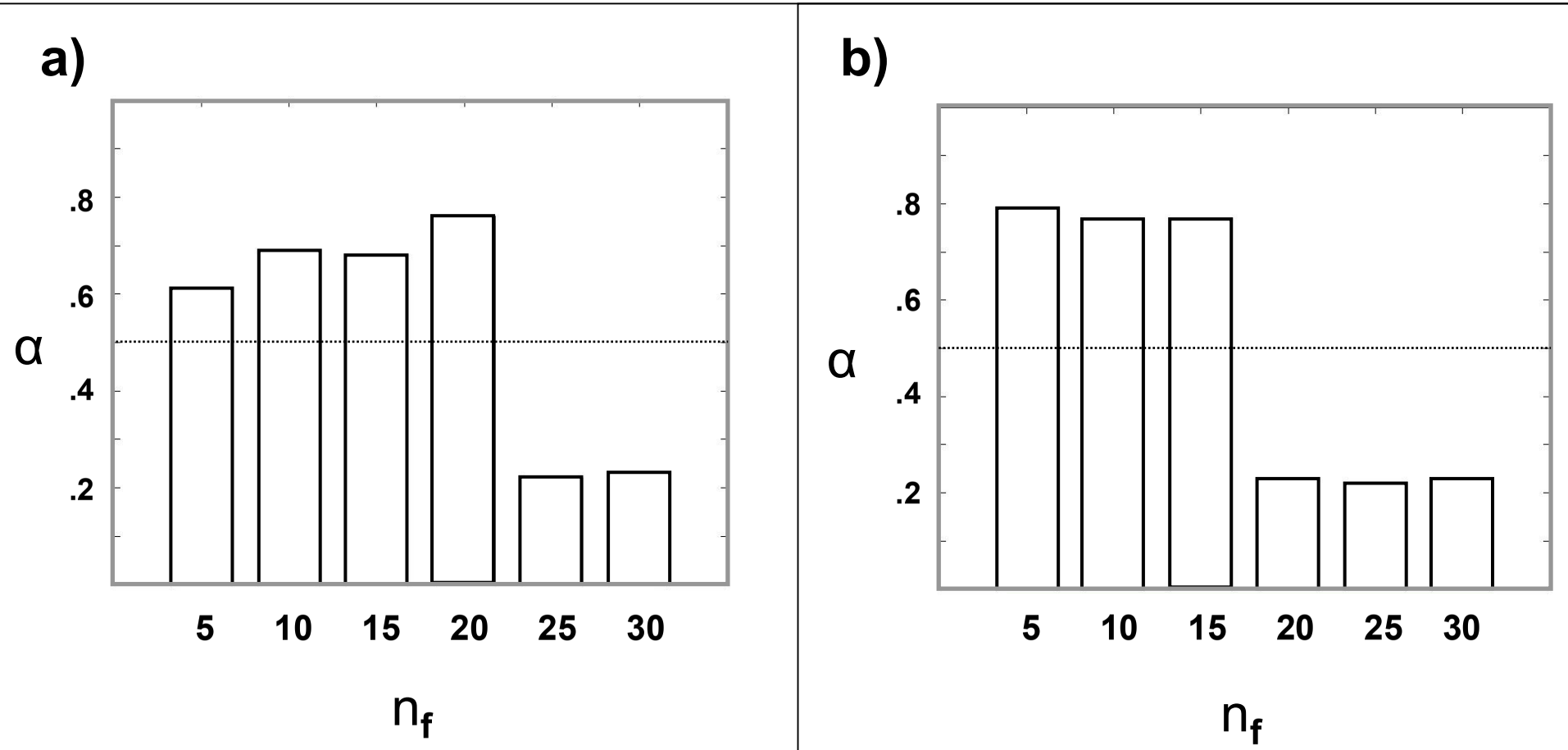

Figure 6. Ensemble joint distribution of 500 hPa geopotential height at two gridpoints. Possible values of height Φ_1 (m) at a gridpoint located at 170.0W longitude, 52.5N latitude define the abscissa, and possible values of height Φ_2 (m) at a gridpoint located at 22.5W longitude, 60.0N latitude define the ordinate. Each point on the plot depicts a particular ensemble member's specific values of height at the two gridpoints. The ensemble members are associated with the 14 May 2003 NCEP GFS 192h ensemble. The grey point corresponds to the control ensemble member, and the black points correspond to the perturbed ensemble members.

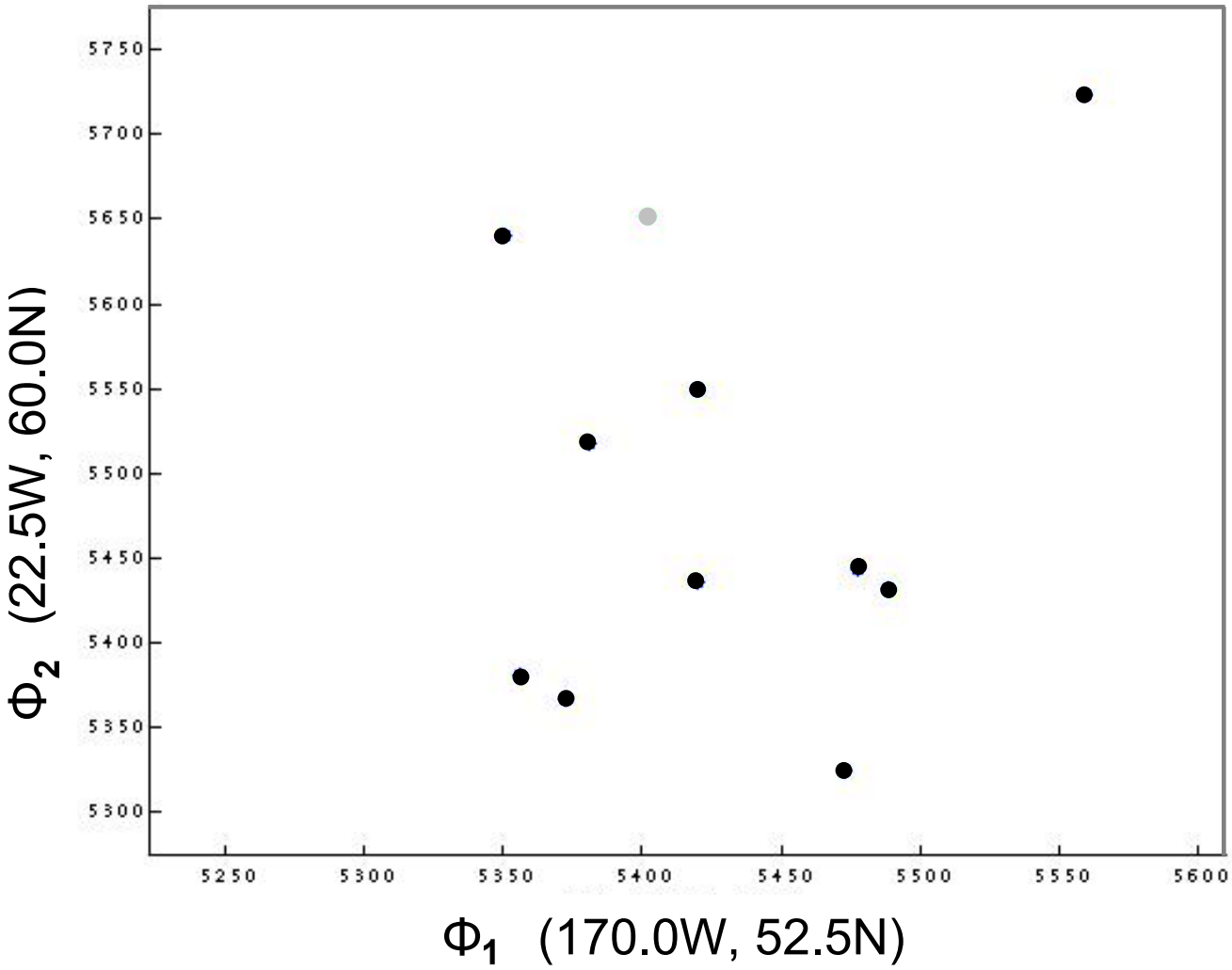

Figure 7. As for Figure 6, except that the dotted lines identify the climatological values of 500 hPa geopotential height $\mathbf{c}_1$ and $\mathbf{c}_2$ at gridpoints 1 and 2, respectively, for 14 May 2003. Also, the labels S_1, S_2, S_3, and S_4 indicate the sample subspaces described in the text.

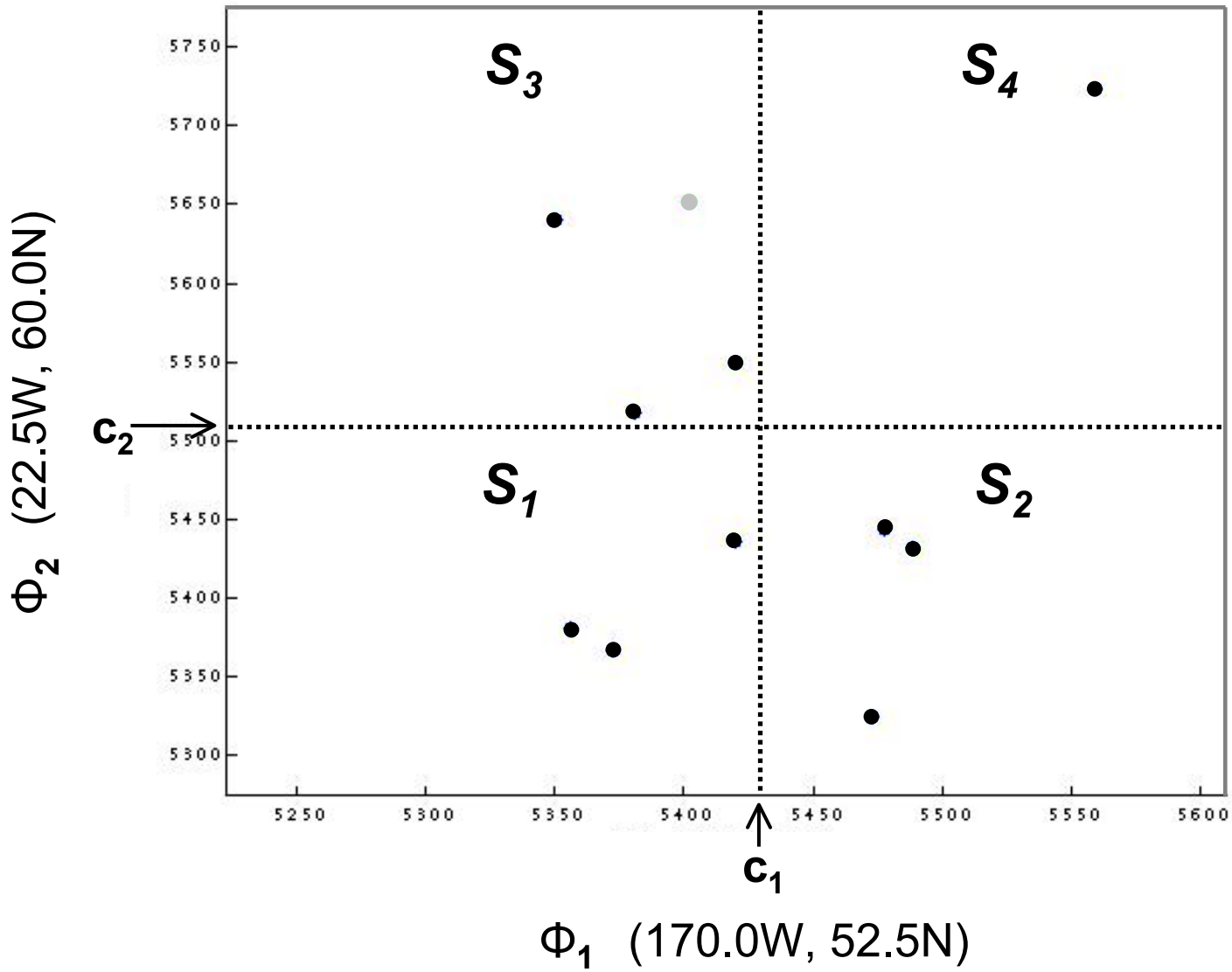

Figure 8. As for Figure 7, except that the pair of signs in each parentheses are the signs of the verifying 500 hPa geopotential height anomalies at gridpoints 1 and 2 in the case that the point corresponding to verification falls into a given sample subspace. The first entry in parentheses corresponds to the sign of the height anomaly at gridpoint 1, and the second entry corresponds to the sign of the height anomaly at gridpoint 2.

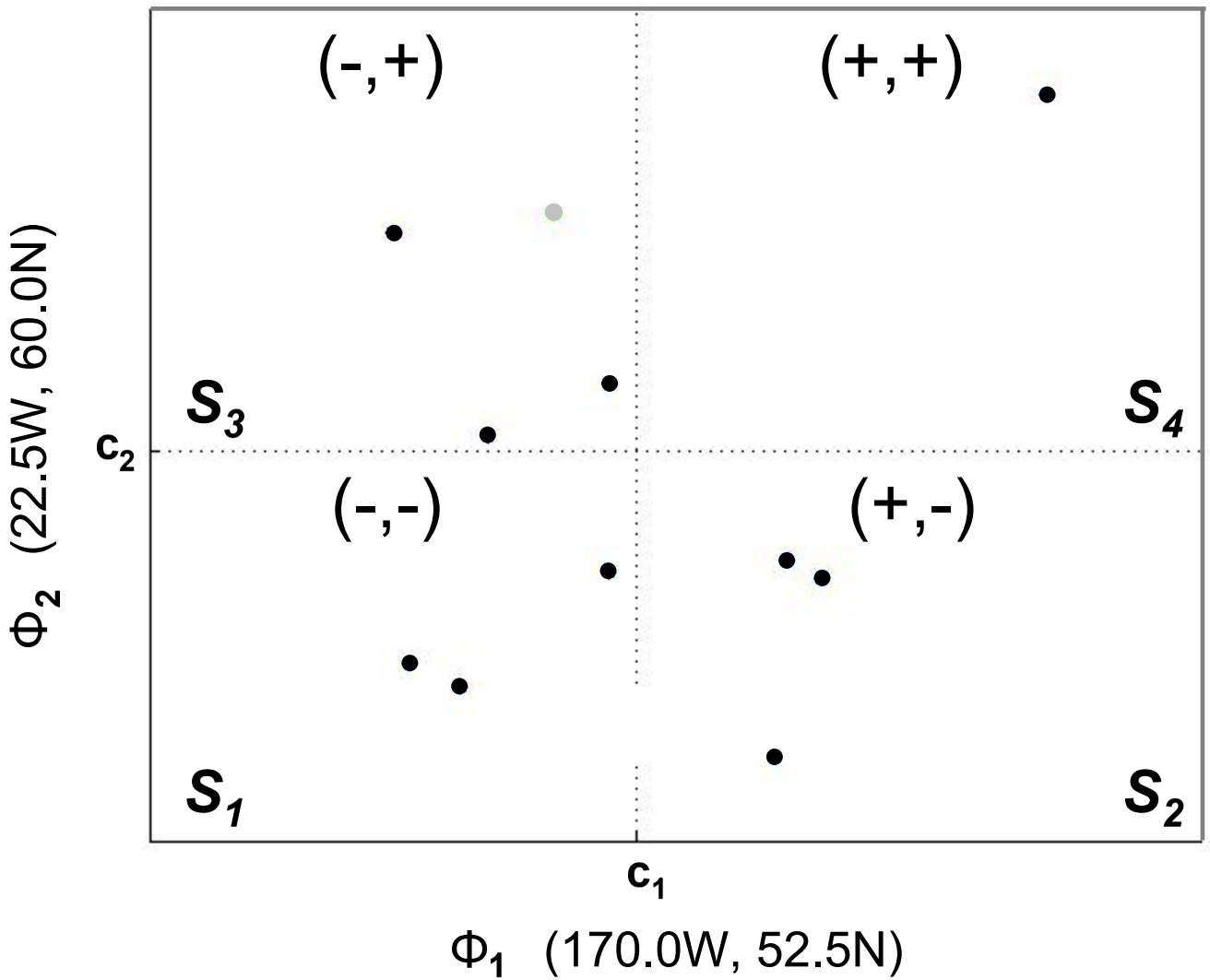

Figure 9. a) Composite 500 hPa geopotential height (m) anomaly pattern associated with the negative phase of the Pacific North American (PNA) pattern. Composite based upon NCEP/NCAR Reanalysis data for the month of January and the years 1968, 1969, 1971, 1972, 1982, and 1989. Contours every 15m. Negative height anomalies ≤ -15m shaded. Figure created using NOAA-CIRES/Climate Diagnostics Center interactive website. b) Horizontal positions of two particular gridpoints located at the 500 hPa vertical level. The position of each gridpoint is indicated by a black dot. The designation of each gridpoint is indicated by the number below the dot.

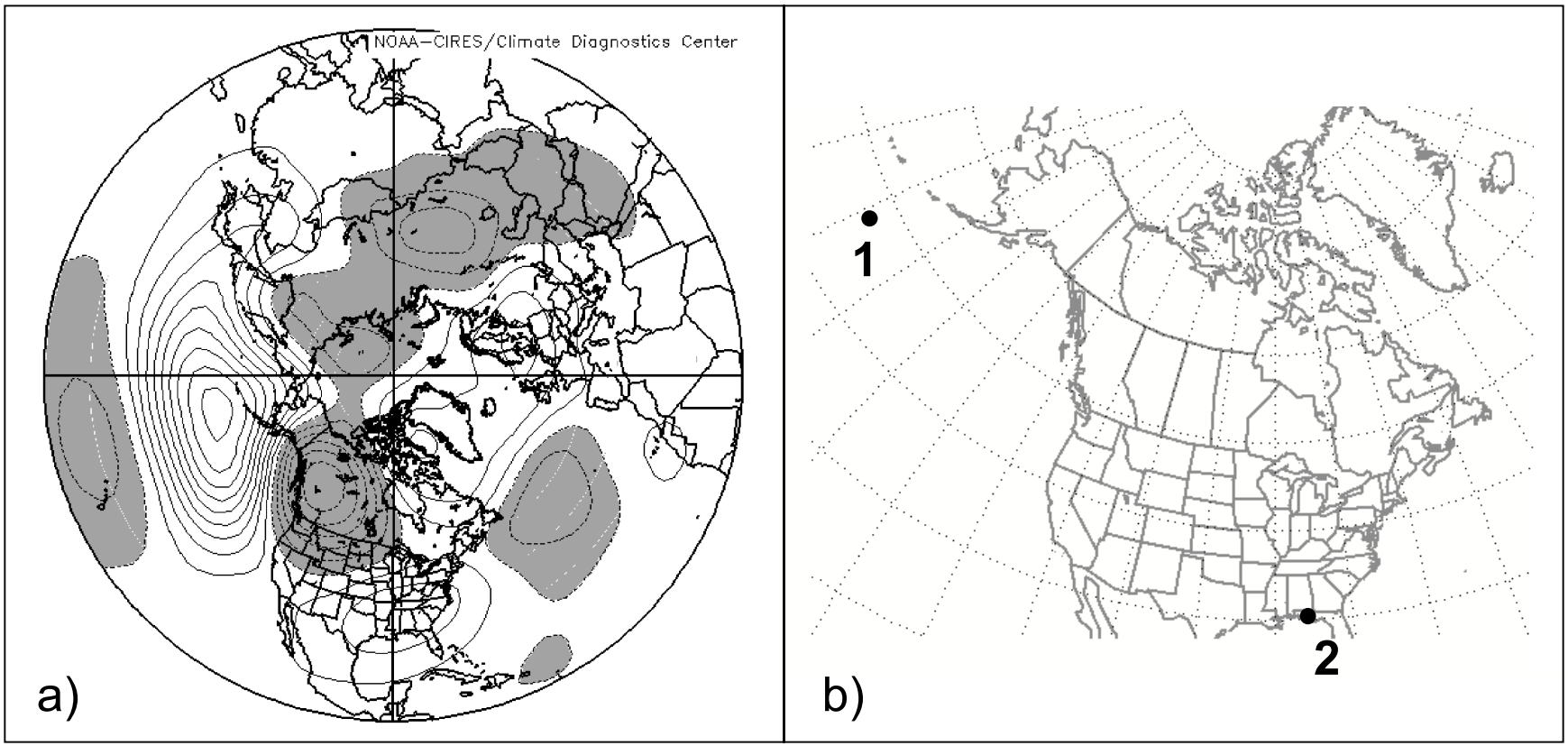

Figure 10. Orientations of the five different gridpoint pairs referenced in the analysis. The horizontal position of a particular gridpoint is indicated by a black dot, and the two gridpoints in a given pair are connected by a black dotted line. The designation of a given gridpoint pair is indicated by the number alongside the black dotted line associated with the given pair.

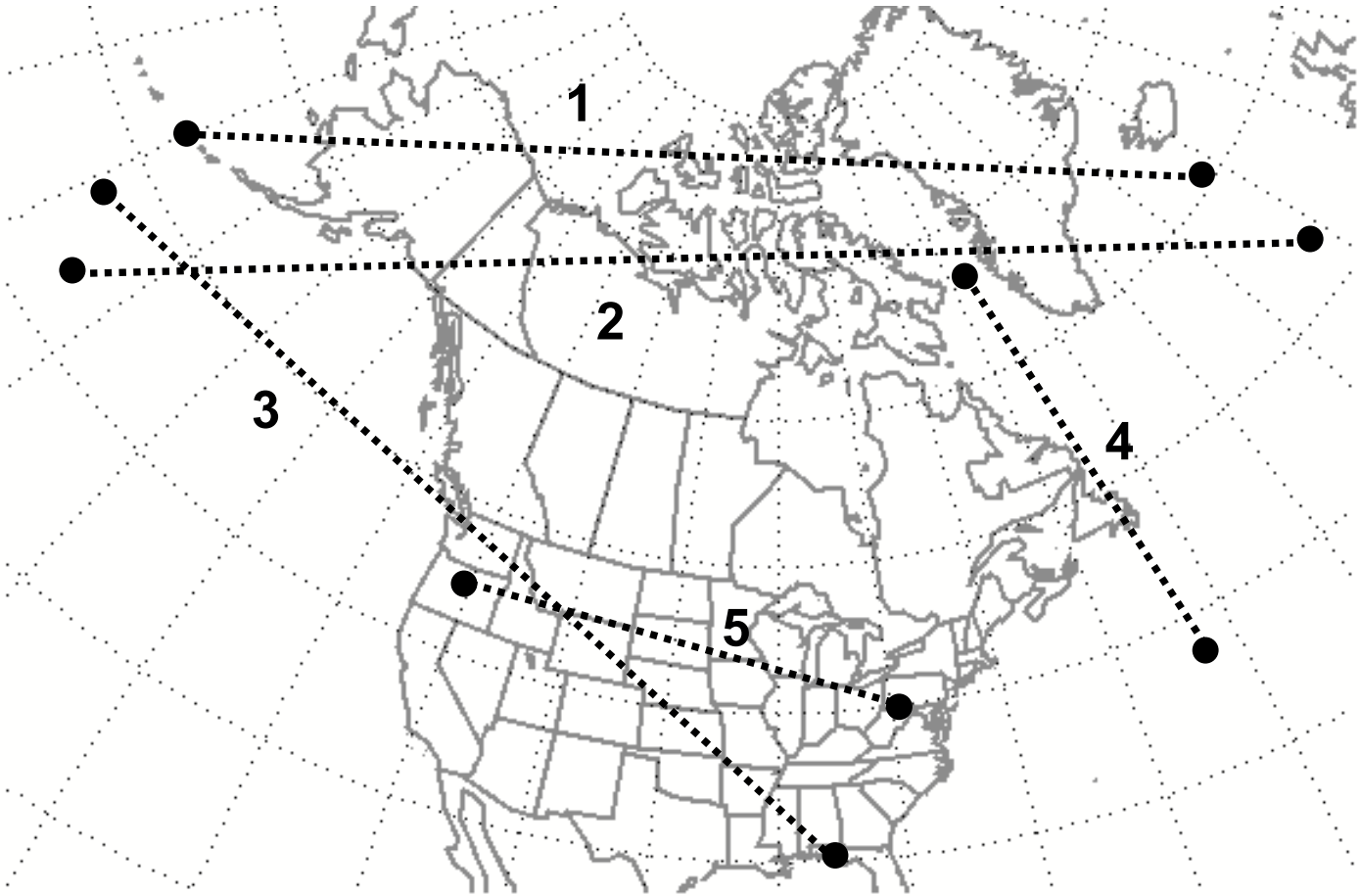

Figure 11. a) Composite 500 hPa geopotential height (m) anomaly field associated with the positive phase of the Pacific North American (PNA) pattern. Composite based upon NCEP/NCAR Reanalysis data for the month of January and the years 1977, 1981, 1983, 1985, 1988, and 1992. Contours every 15m. Negative height anomalies ≤ -15m shaded. b) Composite 500 hPa geopotential height (m) anomaly field associated with the positive phase of the North Atlantic Oscillation (NAO) pattern. Composite based upon NCEP/NCAR Reanalysis data for the month of January and the years 1984, 1986, 1987, 1988, 1989, 1990, 1991, 1993, and 1994. Contours and shading same as for a). Figures created using NOAA-CIRES/Climate Diagnostics Center interactive website.

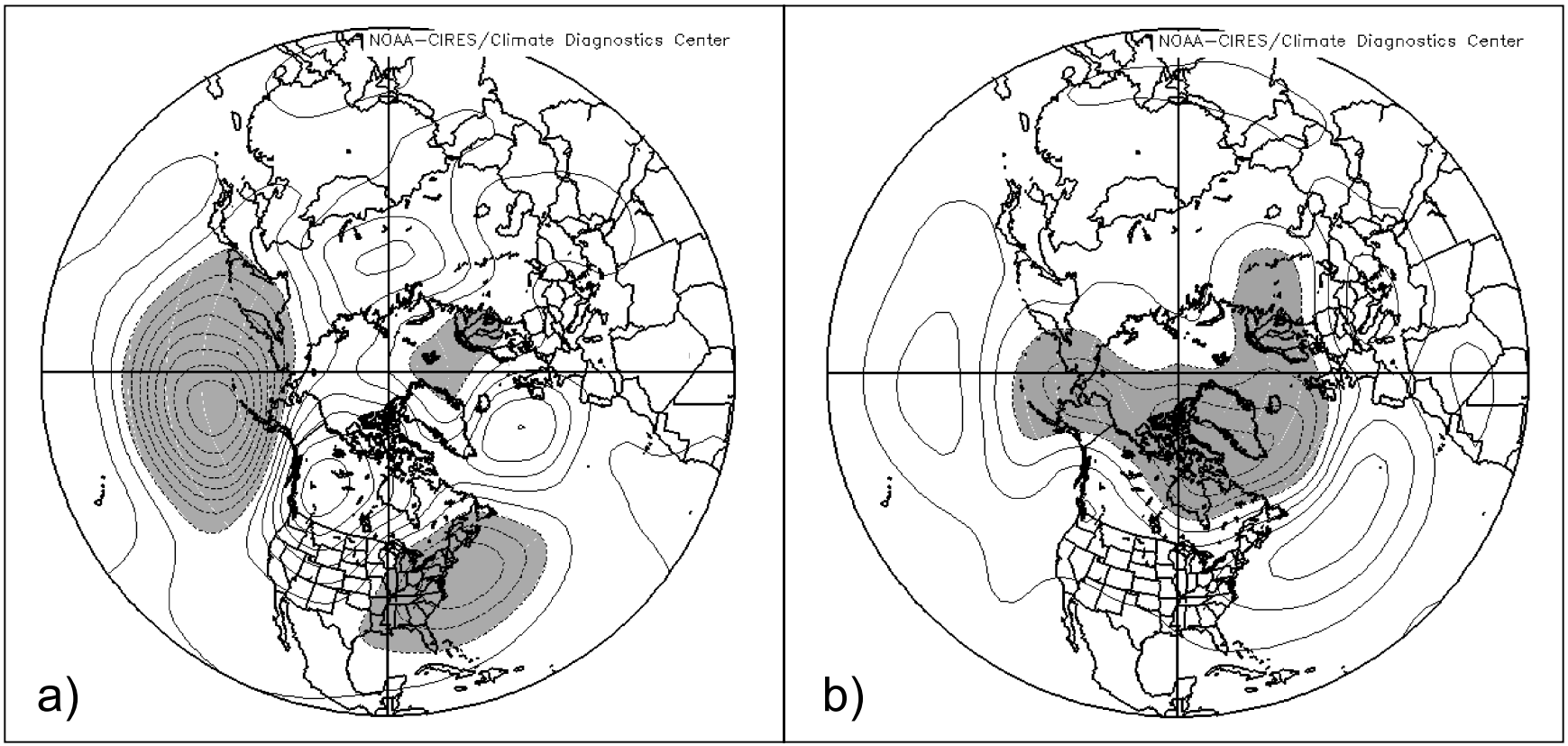

Figure 12. Percentage relative improvement in Brier score derived from hybrid version A.2.17 for gridpoint pair 5 for each of the four dichotomous events (-,-), (+,-), (-,+), and (+,+). The four boxes represent the partitioned sample space (refer to Fig. 8), and the relative improvement is superposed on the sample subspace corresponding to each event. Positive relative improvement less than 5.0% is represented with a check mark.

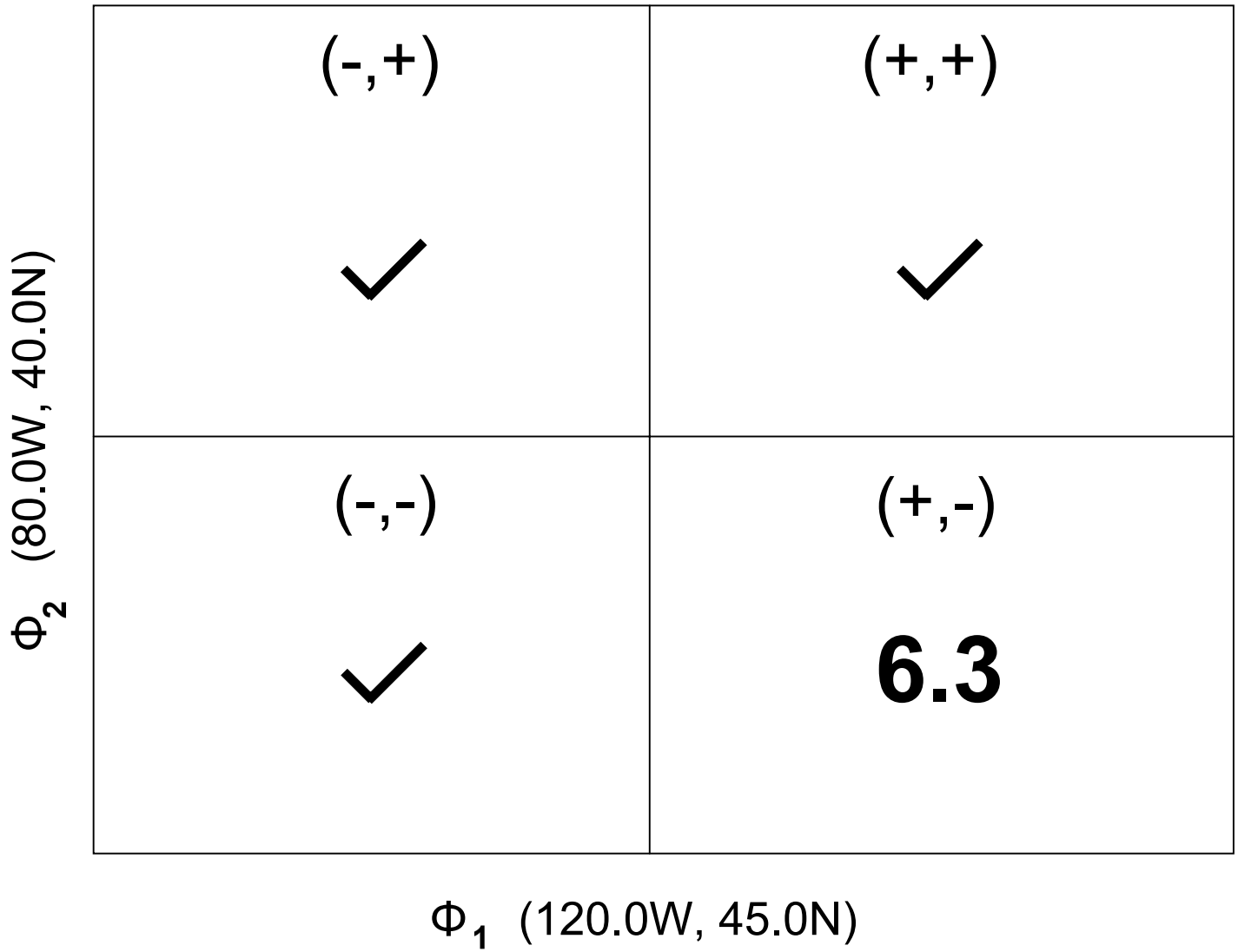

Figure 13. ROC curves of the operational ensemble and a given hybrid ensemble for a particular gridpoint pair and event. The solid (dashed) line is the ROC curve of the operational (given hybrid) ensemble. Each open circle (dot) defines the operational (given hybrid) ensemble's false alarm rate and hit rate for a specific probability threshold. For reference, the dotted line is the ROC curve of an ensemble with zero skill.

a) ROC curves for the operational ensemble and hybrid ensemble version A.2.21, gridpoint pair 1, event (-,+).

b) ROC curves for the operational ensemble and hybrid ensemble version A.1.20, gridpoint pair 3, event (-,+).

c) ROC curves for the operational ensemble and hybrid ensemble version A.1.19, gridpoint pair 4, event (+,+).