Application of the Composite Method to the Spatial Forecast Verification Methods Intercomparison Dataset

JASON E. NACHAMKIN
Naval Research Laboratory, Monterey, California

(Manuscript received 27 October 2008, in final form 3 March 2009)

ABSTRACT

The composite method is applied to verify a series of idealized and real precipitation forecasts as part of the Spatial Forecast Verification Methods Intercomparison Project. The test cases range from simple geometric shapes to high-resolution (~4 km) numerical model precipitation output. The performance of the composite method is described as it is applied to each set of forecasts. In general, the method performed well because it was able to relay information concerning spatial displacement and areal coverage errors. Summary scores derived from the composite means and the individual events displayed relevant information in a condensed form. The composite method also showed an ability to discern performance attributes from high-resolution precipitation forecasts from several competing model configurations, though the results were somewhat limited by the lack of data. Overall, the composite method proved to be most sensitive in revealing systematic displacement errors, while it was less sensitive to systematic model biases.

1. Introduction

With the advent of high-resolution numerical forecasts, repeated experiments have highlighted the limitations of the traditional verification measures (e.g., RMS, threat score) in describing the intrinsic value added by these forecasts (Koch 1985; White et al. 1999; Ebert and McBride 2000; Zepeda-Arce et al. 2000; Colle et al. 2001; Mass et al. 2002; Baldwin et al. 2002; Ahijevych et al. 2009, hereafter AGBE). At issue is the added variance incurred from explicitly resolving finescale phenomena. Even small phase errors can result in major losses in correlation between the forecast and the observations (Murphy 1988). In response, a diverse number of verification methods have been developed to address this problem (e.g., Baldwin et al. 2002; Casati et al. 2004; Ebert and McBride 2000; Davis et al. 2006; Marzban and Sangathe 2006; Roberts and Lean 2008).

Many of the verification methods mentioned above are evaluated in this special issue of *Weather and Forecasting*, which is dedicated to the Spatial Forecast Verification Methods Intercomparison Project (ICP).

This effort was designed to assess and compare the performance attributes of the advanced verification methods. Each method was applied to several sets of idealized forecasts with known errors (both geometric and translational), as well as a set of high-resolution precipitation forecasts. AGBE describe the datasets in detail. The results are meant to help users identify those methods that are most appropriate to their applications. As with any statistical method, users should be aware of the performance attributes before basing any decisions on the results. Recent intercomparisons conducted by Ebert (2008) indicate nontrivial differences may arise between even relatively similar verification strategies.

2. Composite method

Nachamkin (2004) introduced the composite method as a means of verifying oceanic wind forecasts. It is an event-based scheme designed to measure the ability of the model to predict specific weather phenomena. Forecast errors are compiled from the composite mean of the predicted and observed events as a whole (as opposed to calculating averages of errors collected separately from each event). However, when sufficient observations are available, additional statistics can be collected from events individually. This flexibility was chosen in response to data limitations. Many of the events...
in question occurred in remote areas where observations were limited to satellite wind retrievals (Goodberlet et al. 1990). As a result, many wind events were only partially observed, and deterministic verification of any specific forecast was severely limited.

The compositing strategy was patterned after the work of Gray and Frank (1977), which showed the derivation of general hurricane structure fields from sounding composites. Although many meteorological events are not as symmetric as a hurricane, useful information can still be compiled provided the interevent variability is not too great. Instead of the eye, events are composited about the “center of mass” of the homogeneous two-dimensional shape encompassing the event. All points are weighted equally, thus defining the center of mass as the geometric center. Events, as defined by a set of threshold criteria, are first located in the forecasts. Then, the forecasts and any available observations are composited with respect to the center of the predicted events. Like the hurricane composites, large samples produce a mean distribution of the forecasts and observations given that an event is predicted. Likewise, if observed events can be accurately located, a similar composite of forecasts and observations can be generated given that an event is observed. For the oceanic wind events, the model analyses proved to be sufficiently accurate to locate the center of each observed event (Nachamkin 2004). When the observations consist of gridded analyses, as they did in the ICP dataset, the event locations are determined from the observations themselves.

Although the basic premise behind compositing is relatively simple, the event definitions should be carefully considered before going forward. Following the distributions approach of Murphy and Winkler (1987), forecast quality is defined by the similarities between the composite based on the occurrence of observed events and the composite based on the occurrence of predicted events. Variability within the composite should be restricted as much as possible to maintain statistical significance. One should try to composite very similar events together. Constraining the event definitions by size, type, and even shape generally accomplishes this purpose. However, placing too many constraints leads to unrepresentative samples as only small portions of the overall forecast are included in the samples. Nachamkin (2004) and Nachamkin et al. (2005) found broad size categories with no shape constraints produced useful results. Also, the center of the observed events must be defined carefully if the observations are incomplete. Nachamkin (2004) used the model analyses to estimate the centers of the observed surface wind events. However, clouds and precipitation are not well analyzed, and the observed composites must rely on gridded observations. Nachamkin et al. (2005) verified heavy precipitation events, and Nachamkin et al. (2009) verified cloud cover in this manner.

3. Idealized experiments
a. Geometric perturbations

A set of five forecasts consisting of perturbations of an idealized elliptical observed field were specified for verification (AGBE). Each idealized entity was defined by an interior region of high values (25.4 mm) within a larger area of 12.7-mm values. The composite method was applied in each case by defining an event as any feature with a contiguous region of precipitation ≥12.7 mm. The resulting composites were thus centered at the geometric center of the 12.7-mm contour. Since each forecast was verified individually, the composites consisted of one event and were thus very simple. For example, the case I composites (Fig. 1) simply redisplay the geometric shapes. The composite based on the existence of a predicted event (Fig. 1a) is centered on the predicted event, which in this case is the lightly shaded ellipse. The composite based on the existence of an observed event (Fig. 1b) is shifted to the right as it is centered on the observations.

Condensing the general performance over an area into a set of numerical scores can be quite difficult. No single score completely summarizes the complexity of the spatial error, though the conditional bias difference (CBD) has shown some utility (Nachamkin et al. 2005). The CBD is the difference between the mean of the forecast-based and observation-based composite biases over a range of increasing scales as denoted by

\[
CBD_n = \frac{1}{N} \sum_{i,j} N_{FC(i,j)} \sum_{k=1}^{N_{FC(i,j)}} \left( \frac{1}{N_{OC(i,j)}} \sum_{k=1}^{N_{OC(i,j)}} (F_{i,j,k} - O_{i,j,k})_{FC} - (F_{i,j,k} - O_{i,j,k})_{OC} \right),
\]

where the subscripts \( FC \) and \( OC \) represent the composite based on the existence of a predicted and an observed event, \( N \) is the number of samples at point \((i, j)\) on the composite grid, \( n \) represents the horizontal length scale, and \( F \) and \( O \) are the forecasts and observations, respectively. Stated more concisely, \( CBD_n = \text{Bias}(FC_n) - \text{Bias}(OB_n) \), where FC represents the sum at scale \( n \) over the first term in (1) and OB represents the sum over the second term. The calculation is performed over a range

\[^1\] All values within the 12.7-mm contour were weighted equally.
of scales in order to summarize the performance across the composite grid. The calculations are similar to the fuzzy verification fractions Brier score, described by Roberts and Lean (2008), in that they are performed over a series of square boxes of length $n$ (where $n$ is an odd integer) all centered at the central point of the composite grid (marked with an x in Fig. 1). Spatial displacements typically manifest themselves as large errors near the grid center. At larger scales, offsetting near misses and correct negatives lead to reduced errors. Forecasts with small errors or rapid reductions in error with increasing scale are generally considered to be higher in quality.

The case I forecast-based and observation-based mean composite biases and the CBD over the range of box sizes from 1 to 501 points are shown in Fig. 2. Large positive and negative biases at small dimensions reflect the nonoverlapping but closely placed predicted and observed events. The bias of the observation-based composite is negative [$\text{Bias(OB}_n < 0$] due to the prevalence of observed precipitation contained within the smaller boxes centered on the observed event in Fig. 1b. The bias of the forecast-based composite is positive [$\text{Bias(FC}_n > 0$] due to the prevalence of predicted precipitation contained within the smaller boxes centered on the predicted event in Fig. 1a. The CBD is large and positive since $\text{Bias(FC}_n) - \text{Bias(OB}_n > 0$. When events are defined by local maxima, the composite sampling strategy generally results in positive CBD values at small box sizes. As the averaging boxes become larger, $\text{Bias(FC}_n)$ and $\text{Bias(OB}_n)$ trend toward the large-scale bias (zero in this example) as a greater number of offsetting observations and predictions as well as correct negatives are encompassed. The biases converge to zero at $n = 150$, which is the size of the box that fully contains both the predicted and observed events. Small CBD scores are desired (perfect = 0); however, large scores at small box sizes are not necessarily detrimental if the CBD rapidly decreases with box size. Such a situation reflects closely placed forecasts and observations of similar magnitude.

The CBD values for the five geometric cases (Fig. 3) indicate varying degrees of quality. The case I forecast performed the best for all box sizes greater than 100 grid points due to its similarity in shape and close proximity to the observations. Case V scored the best at small box sizes because it was the only forecast to directly overlap the observations. However, case V had the largest (worst) CBD scores at box sizes beyond 250 points due to the displacement of the large precipitation maximum away from the observations. An important caveat of the CBD is that the scores tend to improve with increasing box size due to the inclusion of many correct forecasts of no precipitation. Thus, while the absolute difference between the curves is small, the forecasts can be quite different. Displaying the grid-integrated error at each scale as opposed to the grid-average error can address this issue. Correct forecasts do not contribute to an integrated score, and larger-scale errors are emphasized due to the increased number of points. However, experiments with integrated scoring indicate that it suffers from a lack of sensitivity at small scales. Average scoring was preferred for its portrayal of potentially large errors at small scales. The user should keep in mind that the CBD is a gross aggregation of the error on the composite grid. Alone, the CBD relays little direct information regarding the shape or structure of the error.

FIG. 1. The composites based on the existence of (a) predicted and (b) observed events are displayed for the geometric case I. Forecast values are shaded and observed values are contoured. The center of the composite grid is indicated by the x.
To that end, the composites themselves convey the most information.

b. Field transformations

A second set of idealized experiments was conducted by transforming a gridded precipitation analysis and using the transforms as proxy forecasts for the original field. The analysis grid consisted of the 4-km National Centers for Environmental Prediction (NCEP) g240 grid covering most of the continental United States. On this particular day, convection was active over much of the Midwest and Southeast. The transforms for cases I–V involved simple pattern translations over linear distances, while the transforms in cases VI and VII combined translations with multiplicative and additive data manipulations. The values for case VI were multiplied by a factor of 1.5, while 0.05 in. (1.27 mm) was subtracted from the values in case VII.

In an attempt to investigate the performance of the composite method under extreme conditions, the convective cores were composited. To isolate the cores, an event was defined as any region containing 50 or more contiguous points with precipitation totals greater than 11 mm. This seemingly arbitrary value was a compromise between high rain rates and meaningful results. A value of 12.7 mm (0.5 in.) was originally chosen, but the data contained too few events of this magnitude. Using the 11-mm definition, 8 convective cores (events) were located in cases I–V, 10 events were located in case VI, and 7 events were located in case VII. The convection was extremely variable, and the composites were characterized by relatively small areas of heavy precipitation surrounded by lighter amounts (Fig. 4). However, the translations were clearly discernable in the mean fields. Comparing the CBD values for each experiment (Fig. 5), case I performed the best as might be expected since it was associated with the smallest translations. Error values for cases II–V and VII were relatively similar at small box sizes due to the relatively small size of the convective cores. The case-to-case error growth tended to slow once the diagonal translations exceeded 12 grid points or 48 km, suggesting a limit to the error sensitivity to the CBD. At larger box sizes the CBD traces diverged as the translated precipitation forecasts corrected the composite bias at various distances. Again the CBD was reduced overall at large box sizes due to the effects of correct no-precipitation forecasts. Case VI exhibited the worst performance at small box sizes due to the enhanced bias in the forecasts brought on by the multiplicative factor of 1.5. Cases III and VII were almost indistinguishable except at the smallest box sizes as the bias in case VII was small compared to the translation errors.

To test the sensitivity to event size and magnitude, a second set of composites was taken in which events were defined as regions containing 100 or more contiguous

---

**Fig. 2.** The biases \((F - O)\) from the FC and OB composites from case I (Fig. 1) are displayed. The bias is calculated over a series of expanding square regions centered at the point marked x in Fig. 1. Box size is denoted along the x axis. The CBD represents the difference between the conditional biases (FC – OB) as described in Eq. (1).
grid points with rain amounts greater than or equal to 5 mm. This definition yielded six to eight events in each of the seven experiments. The results were similar to the 11-mm events discussed above except the magnitude of the CBD curves was reduced, and the curves were not as distinct from one another. At the smallest (1 × 1 point) scale, the 5-mm CBD values ranged from 20 to 32 mm for experiments I and VI, respectively. The corresponding

![Conditional Bias Difference](image1)

**Fig. 3.** The CBD values for each geometric case are displayed. Thumbnail diagrams of the predicted and observed shapes in each case are depicted for reference.

![Case IV Composite](image2)

**Fig. 4.** The composite based on the existence of an observed event for the perturbed case IV is displayed. Observed values are contoured and predicted values are shaded at 3-mm intervals. The forecast was created by perturbing the observed field eastward by 24 points and southward by 40 points.
values for the 11-mm events ranged from 31 to 53 mm (Fig. 5). The broader precipitation areas that resulted from compositing the lower rainfall rate events were primarily responsible for the reduction in spread. Though convective precipitation maxima were still embedded within the 5-mm events, these maxima were not directly superimposed. The 5-mm events were also larger in scale, so the probability of overlap was greater for a given translation. At box sizes beyond 25 points, the CBD values for the 11- and 5-mm events converged and approached zero at very similar rates. Since the 11-mm events were subsets of the 5-mm events, the larger boxes contained much of the same information. Also, as noted above, larger scales were more forgiving due to the inclusion of correct negatives and offsetting near misses; thus, the asymptotic behavior is not surprising.

Overall, the most unexpected aspect of the field transformation experiments was the lack of distinction between cases III and VII, and, at larger scales, case VI. The factor of 1.27 mm subtracted from case VII was quite small compared to the rainfall errors associated with the translation. Even when events were defined by reduced rainfall values, the CBD scores remained relatively high due to phase errors in the convective cores. The multiplicative factor of 1.5 in case VI was large enough to register at within the heavier convective cores, but the translation errors dominated at lighter amounts. Since systematic biases are often small compared to errors associated with phase shifts, compositing is probably not the best way to diagnose systematic biases. Fortunately, simple bias calculations are very effective at revealing these errors.

4. Model intercomparison

The final experiment consisted of a series of 1-h precipitation forecasts from three configurations of the Weather Research and Forecast (WRF) model (WRF2CAPS, WRF4NCAR, and WRF4NCEP). This experiment is somewhat different than the idealized experiments because each WRF configuration was evaluated from a series of forecasts as opposed to individual cases. The native grid spacing for WRF4NCAR and WRF4NCEP was 4 km, while it was 2 km in WRF2CAPS; additional details are described in AGBE.

All of the data for the verification study, including the stage II precipitation analyses, were interpolated to the 4-km NCEP g240 grid. Forecasts and observations from nine separate days during the 2005 WRF Spring/Summer High-Resolution Forecast Experiment (Kain et al. 2008) were made available for verification.

To evaluate the performance the three WRF configurations, composites were derived from each configuration and the accompanying observations. Two separate verification studies were conducted using thresholds of 1.27 and 12.7 mm (0.05 and 0.5 in.) to define larger mesoscale events and smaller convective events. Due to the small number of forecasts, the event size ranges

![Fig. 5. The CBD values for the perturbed cases are displayed. The eastward and northward components of the spatial shifts are denoted by the paired values in the legend. In addition to the shifts, the values for case VI were multiplied by a factor of 1.5, while 0.05 in. (1.27 mm) was subtracted from the values in case VII.](image-url)
were liberally defined to include as many samples as possible. These ranges were 100–1000 and 50–1000 contiguous grid points for the 1.27- and 12.7-mm thresholds, respectively. These definitions combined with large precipitation gradients resulted in large variances at each point on the composite grid as events of multiple sizes were sampled and superimposed. Smaller event size ranges allow for less variance; however, since fewer events are sampled, larger sample sizes are required. As mentioned in section 2, the user must consider the compromise between sharp composites and meaningful sample sizes when choosing the event parameters. Most of the convective events contained less than 500 points, but the mesoscale events were well distributed through the specified size range. Some large mesoscale events (not composited) exceeded 3000 points. Over the 9-day sample, 87 and 34 events were located in the observations for the low- and high-precipitation thresholds, respectively. In the WRF2CAPS, WRF4NCAR, and WRF4NCEP forecasts, the low- and high-threshold event counts were (97, 40), (98, 49), and (151, 51), respectively. The models clearly predicted too many events, with WRF4NCEP predicting more events than the other two models.

Very few systematic spatial biases were apparent in the composites. The composite based on the existence of a 1.27-mm event in the forecasts from WRF4NCAR (Fig. 6) typifies the results. Average precipitation from the composited forecasts is focused near the center of Fig. 6a where most of the events overlapped. Away from the center, values were reduced due to the variation in the event size and shape, and the general outward decrease in precipitation intensity. Observed precipitation averages were quite low, and the fields were relatively noisy. The low values do not necessarily mean that every predicted event was a false alarm. If spatial errors are quasi-random, the observed precipitation maxima will be widely distributed relative to the position of the predicted events. As a result, the composite of the observed precipitation given that an event is predicted becomes quite diffuse, and the averages are reduced. Threshold exceedance statistics (described below) indicate that a combination of false alarms and random spatial errors occurred. The lack of statistical significance is reflected in small area over which the forecasts are expected to be greater than the observations by at least 1 mm at the 90% confidence level\(^2\) (Fig. 6b). The low significance was due to the small number of samples with respect to the large standard deviations.

\(^2\) The confidence levels were derived using a \(t\) test with the variances from the observed and predicted rainfall distributions.

**FIG. 6.** The statistics from the composite based on the existence of a 1.27-mm event in the WRF4NCAR forecasts are displayed. (a) Composite mean forecasts are shaded and observations are contoured in mm. (b) The total number of samples is contoured and the region of \(P - O\) differences exceeding 1 mm at the 90% confidence level is shaded. (c) The number of samples exceeding the 1.27-mm threshold in the forecasts (observations) is shaded (contoured).
Although 90 samples may seem adequate, the extreme event-to-event variability resulted in composite standard deviations that were up to 3 times the magnitude of the average values. In situations like this, three or more months of daily forecasts are required for meaningful results. The occurrence frequency of precipitation exceeding 1.27 mm (Fig. 6c) reflected the behavior of the precipitation averages. The composite of the observations was relatively diffuse, though a slight westward displacement of the forecasts was apparent.

Despite the lack of statistically significant spatial bias information, false alarms and missed forecasts can still be tracked by binning results from each separate event. Exceedance frequencies were calculated by counting the number of observed and predicted points that exceeded the two precipitation thresholds within the composite grid. These points were not restricted to the events, and the frequencies were based strictly on the point counts. Results binned by the ratio of the predicted to observed points (Fig. 7) indicate all three model configurations had considerable difficulties predicting the coverage to within a factor of 2. Considering the composite grid was 101 points or 400 km on a side, these errors are quite large. The WRF4NCAR configuration performed best at the 1.27-mm threshold (Fig. 7a), with 49% of the occurrence frequencies falling within a factor of 2. As suggested by the large number of predicted events, the WRF4NCEP stood out for the greatest number of false alarms, with 38% of the forecasts exceeding the observations by greater than a factor of 4. All configurations had even greater difficulties at the 12.7-mm threshold (Fig. 7b) as more than half of the forecasts fell into the factor of 4 bins. WRF2CAPS performed marginally above the others with 29% of the forecasts within a factor of 2 (FAC2) and 54% beyond (above or below) a factor of 4 (FAC4). The WRF4NCAR and WRF4NCEP results were similar with FAC2 values of 27% and 22% and FAC4 values of 53% and 55%, respectively.

The CBD values (Fig. 8) reflect the occurrence frequency statistics to some degree in that they partially depend on intensity. However, the CBD also reflects the ability of each model to predict precipitation in the correct place. In that regard, the curves were quite
similar to one another, indicating no statistically significant differences between the models. At 1.27 mm, the WRF4NCAR had the lowest CBD error values close to the event center, but WRF2CAPS surpassed it at box sizes beyond 17 points. WRF4NCEP had the worst CBD scores at all box sizes at this threshold. The results at 12.7 mm were somewhat surprising in that WRF4NCEP received the best CBD score for all box sizes below 31 points. While WRF4NCEP had the most false alarms, several of its correct forecasts were very close to the event center. The other models tended to be offset to the north and east of the observations, as indicated by the composite based on the existence of an observed event (Fig. 9). The improved 12.7-mm performance of WRF4NCAR in the CBD curves beyond box sizes of 31 points reflects the precipitation region just north of the heaviest observed events (solid contours in Fig. 9).

5. Conclusions

The composite verification method was applied within the context of the ICP to illustrate its capabilities and compare it to different verification methods for spatial forecasts. To that end, a series of idealized and real-data experiments was provided as test data. The composite method performed relatively well in that useful forecast quality information was retrieved, even in situations when very few forecasts were available. This was encouraging because composites are typically applied over large samples covering several months.

The composites of the geometric and perturbed cases were very simple due to the low number of forecast realizations and the idealized prescribed errors. CBD scores generally summarized the performance of each forecast over a range of scales, giving the highest scores to those cases with the smallest perturbations. The CBD was most sensitive to position errors, especially in regions of strong gradients. Although intensity errors did influence the CBD scores, small systematic biases were sometimes overwhelmed by larger errors associated with phase shifts. The magnitude of the CBD tended to decrease with increasing scale due to the large number of correct forecasts of zero precipitation and offsetting near misses. Experiments with integrated scores placed a greater emphasis on the errors at larger scales at the expense of sensitivity at small scales.

The ICP real-data cases were the most challenging due to the extreme variability within the precipitation events. The event composites at both high (12.7 mm) and low (1.27 mm) precipitation thresholds showed modest spatial phase shifts between the forecasts and observations. However, false alarms and large random error components contributed to diffuse composite means. Precipitation threshold exceedance frequencies gathered from the individual events showed that many of the forecast-to-observed ratios were beyond a factor of 2 within a 200-km radius. As might be expected, the performance at the high convective threshold was the
worst, with almost 50% of the forecasts logging ratios beyond a factor of 4, many of those being false alarms. Overall, the WRF4NCAR configuration performed best at 1.27 mm, while WRF4NCEP logged the worst performance with a large number of false alarms. At 12.7 mm, the results were mixed. The WRF2CAPS configuration exhibited the best exceedance ratios, but the WRF4NCEP configuration exhibited the best CBD scores in and near the event center. Since the CBD is more sensitive to phase errors than systematic errors, forecasts with large positive biases may have an advantage if a number of those forecasts are located close to observed events. Notably, the composite of the observed events indicated better spatial performance for those WRF4NCEP forecasts that were not false alarms. The other configurations tended to have slightly greater spatial displacements, though the overall similarities in the CBD scores indicate that the three WRF configurations were not dramatically differently from one another in terms of spatial error.

In practice, composites are best thought of as data-mining applications used to isolate the average conditions associated with well-defined events. Like any data-mining scheme, the parameters need to be chosen carefully to isolate a meaningful sample. Basic constraints on event size ranges often produce good results, but additional constraints can be applied if enough data exist. When compositing very complex fields, such as high-resolution precipitation forecasts, a minimum of 3 months’ worth of daily realizations is necessary for optimal results. The small sample sizes used in the ICP lead to large variances that made the results difficult to interpret. When using the composite scheme, the composite plots themselves are among the most useful output because the spatial distributions serve as a direct assessment of what the user can expect over time. Measures like the CBD are primarily gross summaries of the conditions on the composite grid. Other applications, such as rainfall coverage frequencies and exceedance ratios, are ways to further parse out meaningful trends. The results here indicate that the CBD is best applied to quantify spatial error while systematic biases, false alarms, and missed forecasts are best quantified using exceedance ratios.

Acknowledgments. This research is supported by the Office of Naval Research (ONR) through Program Element N0001408WX21169. The Spatial Forecast...
Verification Methods Intercomparison Project was coordinated by the National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation. Eric Gilleland coordinated the verification intercomparison project and distributed data to the participants. Mike Baldwin, Barbara Casati, and Beth Ebert provided forecast and observation data for the project. The WRF Spring/Summer High-Resolution Forecast Experiment data were provided by the National Severe Storms Laboratory (NSSL), the Storm Prediction Center (SPC), and NCAR. The three anonymous reviewers provided many helpful suggestions that greatly improved this paper.

REFERENCES


